

The
ENCYCLOPEDIA
of
PHYSICS

Edited by

ROBERT M. BESANÇON

*Physical Sciences Administrator
Air Force Materials Laboratory
Wright-Patterson Air Force Base, Ohio*

RETROCONVERTED

B. C. S. C. L.

REFERENCE

894



26 cm.

530.3
B-554

REINHOLD PUBLISHING CORPORATION, New York

Chapman & Hall, Ltd., London

Library of Congress Catalog Card Number: 65-29253
Printed in the United States of America

A

ABERRATIONS

The two sections of this article give a theoretical treatment. For an introductory discussion, see the article entitled LENS.

Geometrical Theory. When a light wave passes through an instrument, the wave front suffers deformations due to the imperfection of the instrument. The optical distance between the emerging wave front and the converging wave front, when the mapping of object points is perfect, is a measure of the aberration of the instrument. It is a function of position of the intersection of the rays with the actual wave front and those with the ideal wave front (perfect mapping). This function is called the *aberration function*. For rotational optical systems, it depends on three invariants, and when the aberration is small, it is expanded in a power series in these invariants. In the following section, we have given the expansion in two different forms, known as the standard and the Zernike-Nijboer expansions and briefly discussed the classification of the terms according to the powers of the invariants. For the analysis of geometrical aberrations and the classification of various types of aberrations, we refer the reader to the references listed in the section below and to the following comprehensive treatise on the subject: Herzberger, M., "Modern Geometrical Optics," New York, Interscience Publishers, 1958.

Diffraction Theory. The starting point of the modern theory of diffraction of optical instruments may be traced to the famous paper on diffraction theory of the phase contrast method by Zernike.¹ The extension to the diffraction theory of aberrations was carried out by him in collaboration with his pupils, especially, Nijboer.² Since then, many advances have taken place, both in theory and experimental observation leading to important applications in the improvement of optical instruments. However, prior to Zernike's pioneering work, some significant contributions to the theory were made by a number of authorities, notably Inगतowski, Fischer, Steward,³ and Picht.⁴

The basis of the diffraction theory of optical instruments is founded on Kirchhoff's integral or a modified form of it, namely

$$U(P) = \frac{ikn}{2\pi} \iint_S \sqrt{KU_0(Q)} \exp ik[W + (\mathbf{r} \cdot \mathbf{s})] dS \quad (1)$$

where $U_0(Q)$ is the value of the field on the wave surface (front) S , K and \mathbf{s} are respectively the gaussian curvature and optical normal vector of S , n is the refractive index of the medium in the image space, and W is the Hamiltonian mixed characteristic of the optical system. The image field $U(P)$ at an image point $P(x, y, z)$ is the geometrical optics wave solution of the scalar wave equation, or Maxwell equations.

A more convenient form of Eq.(1) used frequently in actual problems is

$$U(P) = \frac{ik}{2\pi} \iint_{p^2 + q^2 \leq n^2} g(p, q) \exp ik\phi(P; p, q) dp dq, \\ \phi(P; p, q) = W + (\mathbf{r} \cdot \mathbf{s}) \quad (2)$$

where a point Q on the wave surface $S(Q)$ is represented parametrically in terms of (p, q) , $x = -W_p + \lambda p$, $y = -W_q + \lambda q$, $z = \lambda \sqrt{n^2 - p^2 - q^2}$, $\lambda = \lambda(p, q)$. Here p, q and $(n^2 - p^2 - q^2)^{1/2}$ are the optical direction cosines (components) of the normal vector \mathbf{s} . The amplitude function $g(p, q) = |n| |\Delta|^{1/2} U_0(p, q)$ remains constant along a ray (p, q) in image space, and Δ is the discriminant of the second differential form of S . Equations (1) or (2) are known as the Picht-Luneberg integrals.^{1,5}

The above formulas give all the information about the image produced by an optical instrument for a monochromatic source. Thus the problem is reduced to the evaluation of such integrals over the arbitrary wave front S . In general, W is not known explicitly (closed form) on S , so instead of the wave surface S one takes a reference surface S_0 , usually a spherical wave front with center at the gaussian image point of the optical system, and expands W in a Taylor series in the parameters. In practice S_0 is the aperture (entrance or exit-pupil) of the instrument. For rotational symmetric systems, most frequently employed in practice, W depends on three invariants, $u_1 = x_0^2 + y_0^2$, $u_2 = p^2 + q^2$, $u_3 = x_0 p + y_0 q$, and an additional invariant $u_4 = x_0 q - y_0 p$ for electron optical systems. Therefore, the expansion of W is of the form

$$W = W_0 + \sum a_{1i} u_i + \sum a_{2ij} u_i u_j + \sum a_{3ijk} u_i u_j u_k + \dots, \\ (i, j, k = 1, 2, 3),$$

$$= W_0 + \sum_{p=1}^{\infty} \sum_{n=1}^{\infty} \sum_{m=1}^n b_{lnm} \sigma^{2l+m} \rho^n \cos^m \phi =$$

$$W_0 + \sum_{n=1}^{\infty} \sum_{m=0}^n f_{nm}(\sigma) \rho^n \cos^m \phi \quad (3)$$

l, n, m are positive integers, $n - m$ even ≥ 0 and $u_1 = \sigma^2$, $u_2 = \rho^2$, $u_3 = \sigma \rho \cos \phi$. The individual constants b_{lnm} are the aberration coefficients. This standard development has been used by Steward and others in their treatment of diffraction of aberrations. However, even for individual aberrations of lower order, the evaluation of the diffraction integral leads to complicated expressions for the image field or intensity distribution, since the various orders of a single type aberration are not separated; consequently, it is difficult to separate the contributions of each order to the total intensity in the image plane. For these reasons, both Steward and Picht and later Born obtained incomplete figures of the intensity distributions of the image. On the other hand, Nijboer, following Zernike's ideas, was able to calculate to a great degree of accuracy (unknown before) the intensity distribution for several types of aberrations and high orders. The experimental observations made at Zernike's laboratory, as well as those made at McGill University on microwaves^{2a} are in agreement with Nijboer's figures or with those calculated by Nijboer's method.^{2a}

The Zernike-Nijboer diffraction theory of aberration is based on the development of the aberration function, or W , in terms of orthogonal polynomials (functions) over the region of integration (wave front) which, in their case, was a circle. Instead of Eq. (3), W is expanded in the form

$$W = W_0 + \sum_{n=1}^{\infty} \sum_{m=0}^n f_{nm}(\sigma) Z_n^m(\rho) \cos m\phi,$$

$$(m = 0, 1, \dots, n; n = 1, 2, \dots) \quad (4)$$

where $Z_n^m(\rho)$ are called the Zernike polynomials, which are orthogonal over a unit circle. In this development, a typical aberration term is of the form $b_{lnm} \sigma^{2l+m} Z_n^m(\rho) \cos m\phi$. On account of the orthogonality of Z_n^m , the various orders of a single aberration enter individually (are not mixed) in the expression representing the intensity distribution function; i.e., different aberrations cannot counterbalance each other's contribution for all σ . In general, the amplitude function can also be expanded in Zernike polynomials, or other functions such as Fourier-Bessel, or Dini functions,¹⁰ if the field over the aperture is not constant (coating of lenses). In general, the idea of expanding both the amplitude and the phase function in orthogonal functions over the domain of integration has many advantages over previous methods, since the double integral cannot be reduced into a single integral, except for the simplest type of apertures. However, when only spherical aberration of all orders is considered, the method of integration by parts of the diffraction integral leads to rather simple expression for

the image field. This case has been treated exhaustively for both circular and annular apertures by Boivin.⁸

All the methods discussed above are valid only for small aberrations. For large or moderately large aberrations, one must resort to asymptotic methods, which at present are sufficiently developed to include most of the interesting cases occurring in the theory of diffraction of optical systems. When these analytical methods are combined with the present progress in computational methods, the intensity distribution produced by an optical system can be calculated to any desired degree of accuracy.

NICHOLAS CHAKO

References

1. Zernike, F., *Physica*, **1**, 689 (1934).
2. Nijboer, B. R. A., "The Diffraction Theory of Aberrations," Groningen thesis, 1942. For the experimental part, see the thesis by Nienhuis, K., Groningen, 1948. For microwave experiments see: Bachynski, M. P. and Bekefi, G., *IRE Trans.*, AP-4, No. 3, 412 (1955). "Studies in Microwave Optics," *McGill Univ. Tech. Rept.*, **38** (1957).
3. Steward, G. C., "The Symmetrical Optical System," Cambridge, Cambridge Univ. Press, 1928.
4. Picht, Johannes, "Optische Abbildung," Braunschweig, 1931.
5. Luneberg, R. K., "Mathematical Theory of Optics," Providence, R.I., Brown University, 1944. A new printing of this book will soon be issued by the Univ. of California Press.
6. Linfoot, F. H., "Recent Advances in Optics," London, Oxford Univ. Press, 1955.
7. Born, M., and Wolf, E., "Principles of Optics," New York, Pergamon Press, 1959.
8. Boivin, A., "Théorie et Calcul des Figures de Diffraction de Révolution," Laval Univ. thesis, 1960. It will be published in book form by Laval Univ. Press, Quebec, Canada.
9. Marechal, A., and Francon, M., "Diffraction, Structure des Images," Editions, "Revue d'Optique," Paris, 1960.
10. Francon, M., in "Handbuch der Physik," Vol. 24, Berlin, Springer, 1956.

Cross-references: DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS; LENS; OPTICAL INSTRUMENTS; OPTICS GEOMETRICAL; OPTICS, PHYSICAL.

ABSORPTION SPECTRA

The first experiment in which the light from the sun was dispersed into its spectrum was performed by Sir Isaac Newton in 1666. The chief effect was, obviously, the transformation of the round pin-hole image of the sun in white light into a sausage-shaped array of colors, starting with red and ending, further up the wall, with violet. There was only a gradual transition from one color to the next, and apparently no colors were missing between violet and red.

The next step in the new optical topic of spectroscopy was the observation by W. H. Wollaston (1766–1828) in 1802 that the solar spectrum is not complete, but is crossed by a large number of dark lines—apparently missing wavelengths. A dozen years later Joseph von Fraunhofer (1787–1826) again observed these dark lines in the solar spectrum. In 1859 they were explained by Kirchhoff as due to the fact that the elements which, when in the laboratory, give characteristic bright lines in their spectrum, would in the solar atmosphere absorb those very lines—hence relative darkness is apparent at these places in the spectra when viewed from the earth.

Distinctions should be made at this point between several terms used in the discussion of absorption spectroscopy.

The *absorption coefficient* of a material (α) is expressed in the equation, known as the law of absorption and enunciated by Bouguer and Lambert,*

$$I_x = I_0 e^{-\alpha x}$$

in which the intensity of an incident plane wave I_0 is shown to decrease as the reciprocal of an exponential function to a value I_x after the energy has penetrated to a distance x in the sample of the material. In other words, the fraction dI/I_0 of the initial intensity is "lost" in traversing the distance dx , since $dI/I_0 = -\alpha dx$.

Absorption is the general phenomenon taking place within the body of the material as measured by the absorption coefficient.

Absorbance is the common logarithm of the ratio of the incident to the transmitted intensities.

Absorptance is the measure of the amount of light that disappears at a single reflection.

No substance has been found to exist that does not strongly absorb some wavelengths if the range be sufficiently extended. Dielectrics usually exhibit three extensive regions of large transmission, one in each of the three distinctive portions of the electromagnetic spectrum—very short wavelengths, intermediate wavelengths, and very long wavelengths.

A *blackbody* absorbs all of the radiant energy incident upon it— is a perfect absorber—and likewise acts as a perfect radiator. Kirchhoff's law of radiation states that the ratio of the *emissive power* to the *absorptive power* is the same for all bodies at a specified absolute temperature, or $E/A = \text{a constant} = E_{\text{B}}$. E is the total energy radiated per square centimeter of surface per second and A , the *absorptive power*, is the fraction

* Pierre Bouguer (1698–1758) and Johann Lambert (1728–1777). It was later shown by Beer that the absorption coefficient for a solution is directly proportional to the concentration of the absorbing species. The relationship, known today as Beer's law, is

$$I = I_0 10^{-abc}$$

where a is the *absorptivity*, b is the thickness through which the initial intensity I_0 drops to I , and c is the concentration of the absorbing material.

of the incident energy that is not reflected or transmitted by the surface. Obviously, A is unity for a blackbody, and hence the constant in the above equation is E_{B} , the emissive power of a blackbody at the specified temperature. Absorption lines in a spectrum can be explained on the assumption of *RESONANCE* of the atoms of the absorbing material to that portion of the incident energy spectrum which presents the same oscillation frequency. The atoms reradiate all of the absorbed energy *but in all directions*, so that the portion in the line of sight of the observer is relatively less than what would have been in that position without the intervening vapor.

A material that reduces the intensity of incident light almost entirely without regard to wavelength is said to exhibit *general absorption*. White light becomes gray. In the instances cited in this article, there is *selective absorption*. Flowers, paints, skin, etc., have color by selective absorption since some of the light penetrates ever so slightly into the body of the material.

The absorption bands in the spectra of solids and liquids are usually continuous, gradually fading out along the wavelength axis, but gases show narrow lines in their absorption spectra as a general rule.

We know from elementary optics that the *index of refraction* (n) of a nonconducting material (dielectric) at a definite wavelength is its essential property, for by its use in Snell's law, we can obtain the sequence of deviations of a ray as it passes through or from an interface bounding two media (see REFRACTION). The case is very different for metals (conductors) due to the presence of free electrons in among the atoms. Strong absorption at once occurs so that metals are opaque, certainly to visible light. When the optical properties of metals are being considered, it is more efficacious to use the quantity known as the *absorption index*, defined for a given wavelength λ by

$$\kappa = \frac{\alpha \lambda}{4\pi n}$$

where n is best determined by the measurement of Brewster's angle. (see POLARIZED LIGHT). For silver at $\lambda = 589.3\text{m}\mu$, $n = 0.177$ and $\kappa = 20.554$.

The theory of dispersion shows that generally, in the visible region, transparent materials exhibit a decrease in refractive index with wavelength (section AB of Fig. 1). This part of the graph (AB) is known as the *normal dispersion curve* for the material and can readily be plotted from data taken with a prism spectrometer. The earliest attempt to relate n to λ was made by Cauchy in 1836, namely,

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4}$$

where A , B and C are constants. Although based, as we now know it, on false assumptions, this relationship has proved valuable as a practical working equation, as long as one keeps far from an absorption band. Considering the effect of the

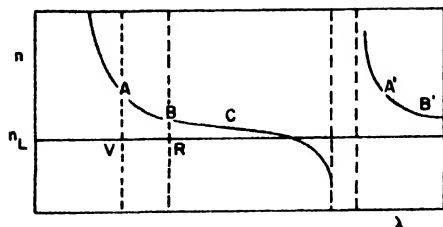


FIG. 1. Index of refraction vs. wavelength for a typical material. Adapted from Fig. 23D of *Fundamentals of Physical Optics*, by Francis A. Jenkins and Harvey E. White, courtesy of McGraw-Hill Book Co., New York, 3rd Edition, 1957.

frequency of the incident light (ν) upon the particles of the medium, having a natural frequency ν_0 , Sellmeier (1871) derived the more acceptable relation

$$n^2 = 1 + \frac{A\lambda^2}{\lambda^2 - \lambda_0^2}$$

where A is a constant proportional to the number of oscillators affected and λ_0 is the wavelength corresponding to ν_0 in a vacuum. We see from this equation that at resonance (when $\nu = \nu_0$), the index of refraction becomes very large. In the event that there are many possible natural frequencies, the Sellmeier equation can be generalized to

$$n^2 = 1 + \sum_i \frac{A_i \lambda^2}{\lambda^2 - \lambda_{0i}^2}$$

A very familiar illustration of absorption is the traffic signal. Even the layman realizes that the practically white light of the high wattage incandescent lamp inside the bowl is *modified* by the presence of the colored material through which the light reaches his eyes. What colors he does *not* see, he assumes are "absorbed." The following experiment may serve to clarify this point.

Let a beam of white light fall upon a flat slab of some dyed material and assume that we are able to measure wavelength by wavelength the percentage of the reflected energy (reflectance R), the percentage of the absorbed energy (absorbance A) and the percentage of the transmitted energy (transmittance T). Because of the law of conservation of energy, it is evident that

$$\begin{aligned} &\text{REFLECTANCE} + \text{ABSORPTANCE} \\ &+ \text{TRANSMITTANCE} = 100 (\%) \end{aligned}$$

(See Fig. 2.)

If the particular values of these characteristic factors are read off at, say, $600 \text{ m}\mu$, it is seen that about 4 per cent is reflected, 31 per cent is absorbed (turned into heat), and the remainder, 65 per cent, is transmitted. The material is said to exhibit *selective* reflectance, absorbance and transmittance, since not all wavelengths are equally affected. In fact, it is observed from the

curves that the slab is probably somewhat reddish by reflected light and amber by transmitted light. No light at all is passed below about $520 \text{ m}\mu$.*

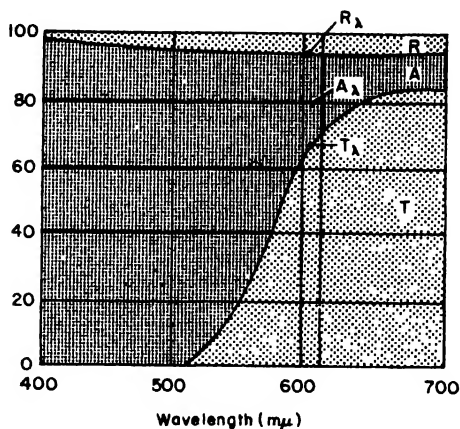


FIG. 2. Distribution of energy vs. wavelength for a slab of dyed material.

The situation would be somewhat different if the sample of material was a thin film of a normally colorless material, such as glass, gelatin, cellophane or acetate, which has been colored by means of a dye—as in the case of a filter. The first surface of the vehicle will not exhibit any selective reflection and the energy distribution will be that of the light source. (It should be observed, however, that the *amount of energy reflected* will depend upon the *angle of incidence*, in accord with Fresnel's reflection formulas.) The dyed material will show selective absorption and transmission. The transmittance curve for a certain blue filter, obtained with a spectrophotometer, is given in Fig. 3. Note should be made of the broad absorption band between about 550 and $630 \text{ m}\mu$. Still more striking is the spectrophotometric curve for didymium (Fig. 4), a material shown originally by von Welsbach to be a mixture of two rare earth elements, praseodymium and neodymium, each of which in solution exhibits intense absorption spectra. To the eye, a didymium filter appears to be pale pink. The range of absorption becomes sharper as the temperature approaches that of liquid air.

Absorption in the visible region (400 to $700 \text{ m}\mu$) has been illustrated by the transmittance curve of the blue filter (Fig. 3) and may be quite graphically understood by the behavior of many dyes in the solid state. As an example, fuchsin exhibits strong absorption in the green region of the spectrum and at the same time reflects a brilliant green. As an absorption band is approached during the determination of dispersion data (Fig. 1), the

* The hue ranges in the visible spectrum are ($\text{m}\mu$): violet, 400 – 450 ; blue, 450 – 500 ; green, 500 – 570 ; yellow, 570 – 590 ; orange, 590 – 610 , and red, 610 – 700 .

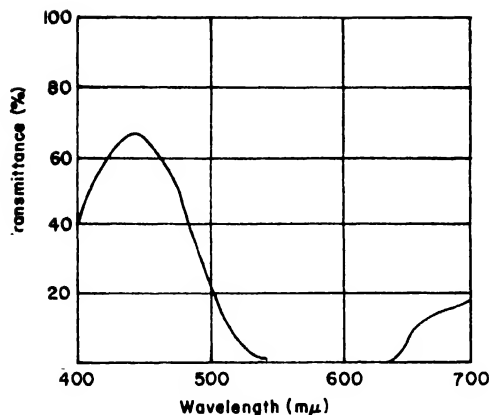


FIG. 3. Transmittance curve for a typical blue filter.

index of refraction is found to change rapidly as the band is entered; in fact, measurements become very difficult if not impossible. Reflection is greater in this region than elsewhere. The matter is summed up in the first diagram (Fig. 1) indicating that the *normal dispersion* curve for the material extends from A to B, where the Cauchy equation holds quite well, but behaving in a peculiar manner at some point C where the Cauchy equation will not give the observed effects. Once the absorption band is passed, as indicated

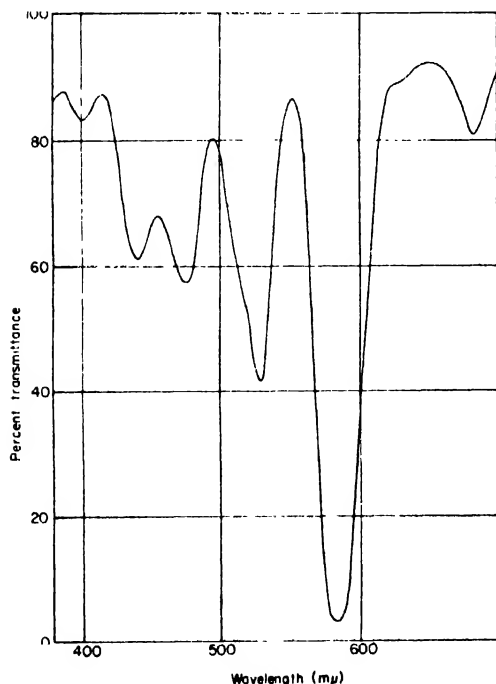


FIG. 4. Transmittance of a didymium filter.

by the two vertical dashed lines, in this case in the infrared, Cauchy's equation, with new constants, begins again to apply, A'B' corresponding to AB. Absorption bands can be observed for hydrogen chloride in the infrared and for ammonia in the microwave region.

X-ray absorption spectra form a very important part of the subject of absorption. Essentially the procedure is to direct a narrow beam of x-rays upon a sample of the material (of thickness x), and then allow the modified beam to impinge upon the atom planes of a crystal of known spacing (d). The diffracted x-rays come out of the crystal at such angles θ as to satisfy Bragg's law

$$n\lambda = 2d \sin \theta$$

and are received by a detector. The latter may be an ionization chamber or a photographic plate. Each setting of the x-ray spectrometer corresponds to a definite wavelength. The values of the intensities measured by the detector with the sample in place (I) and without the sample (I_0) are found from

$$I = I_0 e^{-\mu x}$$

where μ is the absorption coefficient of the material. If we are to take account of scattering and possible fluorescence, then we must write

$$\mu = \sigma + \tau$$

where μ is now the total fraction of the incident energy lost by the beam per unit volume of material, σ is the fraction of the incident energy scattered per unit volume and τ is the fraction transformed into *fluorescent radiation*. If ρ is the density of the material, then the mass absorption coefficient* of the material is

$$\frac{\mu}{\rho} = \frac{\sigma}{\rho} + \frac{\tau}{\rho}$$

and represents the fraction of the energy removed from the incident beam of unit cross section per unit mass. If M is the atomic mass of the isotope of the element being investigated and N_0 is Avogadro's number, then the *atomic absorption coefficient* is defined by

$$\mu_a = \frac{\mu}{\rho} \frac{M}{N_0}$$

As an example, for copper, with x-rays of wavelength 7 mμ, the value of μ_a for copper is 50×10^{-22} cm²/atom. From a study of the variation of μ/ρ with λ , knowledge of the energy levels within an atom has been gained.

Included among the practical uses of absorption spectroscopy as an experimental tool are: (a) to learn which wavelengths of electromagnetic radiation are absorbed, (b) to discover how much of this energy is absorbed under given conditions and then, (c) as a final goal, to find out *why* absorption takes place—and *where*. If we wish to obtain information as to the location and

* Considerable research on mass absorption coefficients has been done by Professor S. J. M. Allen at the University of Cincinnati.

brightness of the absorption lines or bands, a spectroscope or spectrograph is used; if we wish to learn how the energy is distributed throughout the spectral region under investigation, we use a spectrophotometer. Reflectance and transmittance curves are obtained with the latter instrument and a specification of the "color" of the sample in either case can be determined using the C.I.E. methods.

Absorption spectrophotometry is the technique whereby one establishes the relationship between wavelength and radiant energy as the latter proceeds into a given material.

C. HARRISON DWIGHT

References

- Bauman, Robert P., "Absorption Spectroscopy," New York, John Wiley & Sons, 1962.
 Harrison, George R., Lord, Richard C. and Loofbourow, John R., "Practical Spectroscopy," Englewood Cliffs, N.J., Prentice-Hall, 1948.
 Evans, Ralph M., "An Introduction to Color," New York, John Wiley & Sons, 1948.
 Jenkins, Francis A., and White, Harvey E., "Fundamentals of Physical Optics," Third edition, New York, McGraw-Hill Book Co. 1957.

Cross-references: COLOR; OPTICS, PHYSICAL; POLARIZED LIGHT; RADIATION, THERMAL; REFRACTION; SPECTROSCOPY; X-RAYS.

ACCELERATION. See DYNAMICS.

ACCELERATORS, LINEAR

Linear accelerators (often abbreviated to "linacs") are used for acceleration of electrons, protons and heavy ions. Electron linear accelerators have yielded electrons at energies above 1 BeV; proton linear accelerators have not yet reached energies above 70 MeV.

Although the term "linear accelerator" is occasionally used to describe systems in which particles are accelerated by electrostatic fields (Cockcroft-Walton or electrostatic accelerators), the term is generally used to apply to systems in which particles are accelerated along a linear path by application of rf fields. Only accelerators of this type will be discussed in this article.

The linear accelerator has the advantage that the accelerated beam is easily extracted for experimental use. In principle it is capable of producing well-focused beams of higher intensity than are available from circular machines of the synchrotron or synchrocyclotron type. It does, however, require very high power levels at frequencies where conversion equipment is relatively expensive. For a given final energy, a linear accelerator will usually be materially more expensive than a synchrotron. (For a general discussion of accelerators see ACCELERATORS, PARTICLE.)

Field Patterns used in Linear Accelerators. The rf fields used for acceleration are set up in a long cylindrical cavity whose axis is to be the axis of the accelerated beam. Hence for acceleration the field pattern must have a major electric field component parallel to the axis. This requirement is satisfied by the TM_{01} waveguide mode in which a paraxial electric field has its maximum strength at the axis and falls to zero at the cavity wall. Azimuthal magnetic fields lie in planes normal to the axis, have small values near the axis and increase to maximum values at the cavity walls. Usually the field pattern is maintained by coupling to these magnetic fields by loops or apertures excited by external power sources. Corresponding to the high rf magnetic field at the wall, paraxial currents flow in the walls and are responsible for a major fraction of the power loss in the system. When high electric fields are required on the axis to accelerate to high energy in reasonable distances, the wall currents are correspondingly high. For acceleration rates of 2 MeV/m, power losses in copper walls will be of the order of 50 kW/m.

Both standing wave and traveling wave patterns are used in linear accelerators. If traveling waves are used, as is the case in most *electron* machines, the phase velocity of the waves must be made equal to the velocity of the particles accelerated; as the particle velocity increases, the phase velocity also must increase. But phase velocities in simple waveguides always are greater than the velocity of light, and loading must be introduced to reduce the phase velocity to the desired value. This is accomplished by introduction at intervals of washer-shaped irises, as shown in Fig. 1.

Standing wave patterns are used in *proton* linear accelerators. Cavities many meters in length

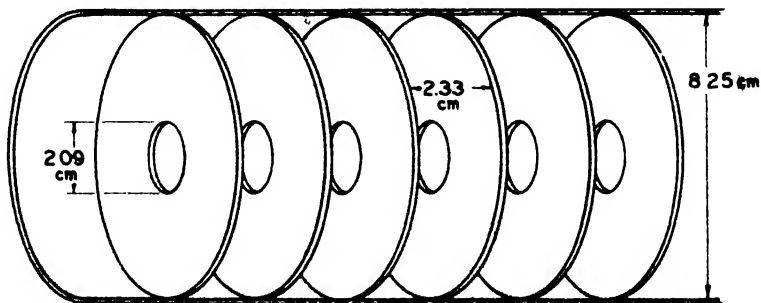


Fig. 1. Cutaway of iris-loaded waveguide for electron linear accelerator.

are excited in the TM_{010} mode in which the axial field is uniform from one end of the cavity to the other. Protons which enter the cavity at a low injection velocity may arrive at a phase of the rf field at which they are immediately accelerated, but before they have traveled more than a few centimeters, the field will reverse and become decelerating. To protect the particles from the field in its decelerating phase, "drift tubes" are introduced, as shown in Fig. 2. These are pipes

Losses per unit length in waveguides generally decrease as the square root of the rf wavelength for equal axial fields. Hence, where possible, it is desirable to operate at as high a frequency as possible. But, as wavelength is decreased the diameter of the structure and of the beam aperture decrease correspondingly. The highest frequency that gives convenient beam apertures and at which adequate power sources are available is in the 3000-Mc/sec range. For reasons that are pri-

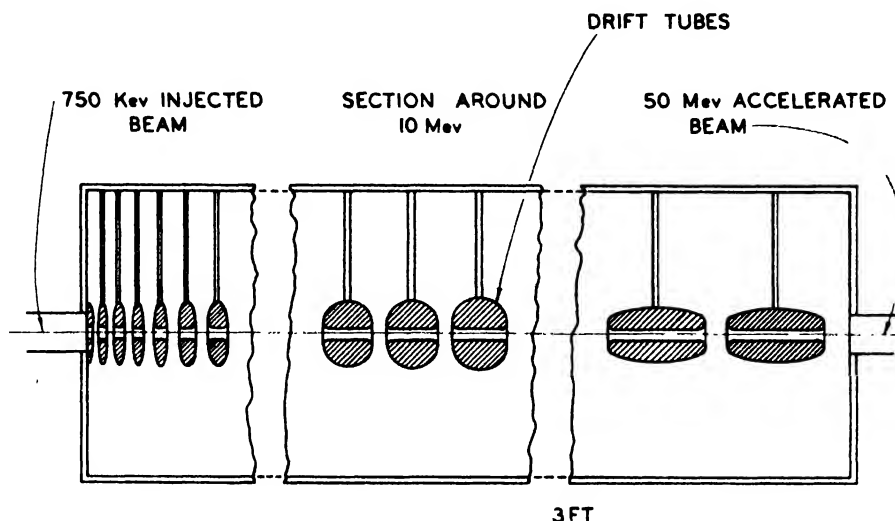


FIG. 2. Cross section through proton linac.

coaxial with the cavity and of such length that the particle is protected from the field during its reverse phase and emerges only after a complete rf cycle when the field again is accelerating. As the particles gain energy, the drift tubes are increased in length.

It would appear that the rather complicated drift-tube structure is conceptually and mechanically inferior to the rather simple iris-loaded traveling-wave system. It is adopted at the relatively low phase velocities required for protons in the range below about 200 MeV because the extreme loading required to reduce the phase velocity of the iris-loaded system to velocities below one-half of the velocity of light results in very high losses. From the point of view of rf power consumption, the drift-tube structure is much superior at low phase velocities.

Electron Linear Accelerators. Electrons very rapidly approach the velocity of light (c) as they are accelerated. At 1 MeV an electron already has reached 94 per cent of its ultimate velocity. At energies higher than this satisfactory acceleration will be achieved if all sections of the accelerator are made to have phase velocities equal to c . This makes much easier the tasks of construction and of operation. For example, rf excitation of a section of the accelerator may fail and the whole machine will still be operative, although at a slightly lower final energy.

marily historic, most electron linear accelerators in the United States are operated at a frequency of 2856 Mc/sec.

Both the phase velocity and the group velocity in the guide are determined by the dimensions of the guide and the loading irises. The group velocity is fixed also by the capabilities of the rf power sources. Klystrons with outputs of the order of 20 MW have become standard; each klystron can excite a section of waveguide 3 meters long to axial fields of 10 MV/m. The group velocity suitable for this operation is 1 per cent of the velocity of light. The dimensions indicated in Fig. 1 result in a phase velocity of c and a group velocity of $0.01c$ when the guide is excited at 2856 Mc/sec.

Injection is from a conventional electron gun. In some cases a short "bunching section" pre-groups the electrons around the peak of the accelerating wave. In this section, the phase velocity is matched to the electron velocity by suitable choices of dimensions.

The power levels required are so high as to preclude continuous operation. Typical operation is with two-microsecond ($2\text{-}\mu\text{ sec}$) pulses repeated several hundred times per second. Of the $2\text{-}\mu\text{ sec}$ pulse, the first half is required to build up the accelerating field.

Electron linear accelerators in the energy range below 100 MeV are widely used for x-ray

production and are commercially available. Most of the pioneer work on electron linear accelerators was done at Stanford where a machine is now (1964) operating at 1.2 BeV and another, two miles long, is under construction and is aimed at eventual operation at 40 BeV. At Orsay, France, a 1.3-BeV electron machine is in operation, and in Kharkov in the U.S.S.R. a 2-BeV accelerator is expected to be in operation during 1964.

Proton Linear Accelerators. Because of the lower velocities of protons at million-electron-volt energies, proton linear accelerators suffer from several limitations from which the electron machines are free. Injectors for protons usually are Cockcroft-Walton voltage multiplier sets giving energies of 500 to 750 keV. At 750 keV the velocity of a proton is only 0.04c. The accelerating field component at such low phase velocities varies strongly with radius at a rate that is approximately proportional to the square of the frequency. This effect sets an upper limit of about 200 Mc/sec for the frequency of the accelerating field, and most proton linear accelerators are operated in the neighborhood of 200 Mc/sec. A cavity resonant in the TM_{010} mode at that frequency will be about 90 cm in diameter.

Figure 2 is a schematic cross section through a 50-MeV proton linac used at the Brookhaven National Laboratory as the injector for the 33-BeV synchrotron. Sections are shown at the injector end, at the region where the protons have an energy of about 10 MeV, and at the high-energy end. The over-all length of the machine is about 33 meters. The drift tube shapes indicated have the purpose of keeping each section of the machine resonant to give a uniform accelerating field pattern and, at the same time, of holding the resistive losses in the walls of the drift tubes to levels as low as possible.

The principle of phase stability is operative in proton linacs whereas, at the extreme relativistic velocities of multi-MeV electrons, electron linacs do not enjoy phase stability and require extreme precision in axial dimensions. In the proton linac, the drift tube lengths increase at a rate corresponding to acceleration at a phase displaced 20 or 30° from the peak of the wave. The phenomenon of phase stability (see ACCELERATORS, PARTICLE) results in continual restoration to the correct phase of protons which enter the machine at phases in the neighborhood of the correct phase. Often prebunchers are used to collect a large fraction of the injected beam around the accelerating phase. These prebunchers have the same design as the modulating gap in a klystron and function in the same fashion.

At the stable phase the field across an accelerating gap is rising as the proton crosses the gap. As the proton enters the gap, the accelerating field has a focusing component, but as it enters the next drift tube it feels a larger defocusing field and the net effect is a strong defocusing. In early proton linacs this effect was overcome by the introduction of rudimentary grids at the downstream end of each gap. These grids give unsatisfactory performance because they intercept a

large fraction of the beam and because their poor optical quality results in loss of many protons. With the advent of alternating gradient focusing, grids in linacs were largely abandoned and focusing is now accomplished by quadrupole magnets imbedded in the drift tubes. This has resulted in an increase in output current by two orders of magnitude to levels of the order of 50 mA.

As in electron machines, the high rf power level required forces operation at a relatively low duty cycle. Since, at this frequency, the time required to build up the field in the linac cavity is about 200 μ sec, pulse lengths for research use are chosen to be several hundred microseconds. Duty cycles are rarely larger than 1 per cent.

The first proton linac was the 40-foot machine at the Lawrence Radiation Laboratory in which protons were accelerated to 32 MeV. Fifty-MeV linacs are used as injectors for the synchrotrons at Brookhaven, Argonne and CERN. The largest linac now in operation is the 70-MeV unit in research service at the University of Minnesota, but a 100-MeV machine is under construction in the U.S.S.R. to serve as the injector for the 70-BeV synchrotron at Serpukhov. At CERN and in the United States, there is, in 1965, much discussion of proton linacs for energies as high as 800 McV for use as synchrotron injectors or as meson research facilities.

Heavy-ion Linear Accelerators. Heavy ions such as C, N, O, Ne and even higher masses can be accelerated in a structure similar to a proton linac but operated at a lower frequency, typically 70 Mc/sec. Multiply charged ions are injected at a few hundred kiloelectron volts and are accelerated to about 1 MeV/nucleon. They then pass through a "stripper," a gas jet in which their charge-to-mass ratio is doubled. The accelerating field now can be twice as effective, and the ions enter a second cavity in which they are accelerated to about 10 MeV/nucleon.

Heavy ion linacs are used in research in nuclear physics and nuclear chemistry and have been particularly useful in the production of transuranium elements.

JOHN P. BLEWETT

References

- Smith, Lloyd, "Linear Accelerators," in "Handbuch der Physik," Vol. 44, pp. 341-389, Berlin, Springer-Verlag.
- Livingston, M. S., and Blewett, J. P., "Particle Accelerators," Ch. 10, New York, McGraw-Hill Book Co., 1962.
- Livingood, J. J., "Principles of Cyclic Particle Accelerators," Ch. 14, Princeton, N.J., D. Van Nostrand, 1961.

Cross-references: ACCELERATORS, PARTICLE; ACCELERATORS, VAN DE GRAAFF; BETATRON; CYCLOTRON; SYNCHROTRONS.

ACCELERATORS, PARTICLE

The primary purpose of particle accelerators is to serve as artificial sources of well-defined beams of radiation for studies in nuclear science. Not only are the intensities much greater than those provided by cosmic rays, but the energies and directions of motion of the particles are known to high precision. In addition, several types of accelerators are useful for providing x-rays and beams of neutrons for medical applications. Industrial uses include the production of radioactive-tracer materials and beams for sterilization processes.

The basic components of an accelerator consist of a source of particles, an evacuated chamber to keep them from being scattered by air molecules, and a means of providing them with kinetic energy. Acceleration is always accomplished by means of an electric field since mechanical forces or gravitational fields are too weak and magnetic fields only produce a change in direction of the particles. The various types of particle accelerators merely reflect the different methods of applying the electric fields. Only charged particles can be accelerated, but these may include electrons, protons, deuterons, or heavier ions such as multiply-ionized carbon, nitrogen, etc. Beams of neutrons are obtained by bombarding suitable targets (D, Li, Be, etc.) usually with deuterons.

The simplest type of electric field is provided by applying the required amount of voltage between two terminals. About one-half of all existing accelerators are of this DC (direct potential drop) type, but their energies are limited to a few million electron volts.*

The methods by which this direct voltage may be obtained are of two basic types: (1) by means of a cascade process such as the cascade rectifiers or voltage-multiplying circuits and (2) by charging up a terminal through actual transportation of charge.

The *Cockcroft-Walton accelerator* belongs to the former type and consists of several stages of a voltage-doubling circuit together with an ion source and a suitably designed discharge tube. Although it is possible to use electrons, these accelerators are usually used as positive-ion sources and can provide dc currents up to about 10 mA and energies up to about 1.5 MeV without special pressure tanks.

The second method is employed in the *electrostatic*, or *Van de Graaff*, generator where a row or corona points sprays charge onto a moving belt that carries the charge to the field-free region inside a spherical metal terminal. Currents of about 1 mA for electrons, and up to about 500 μ A for positive ions, are obtained in the range 1 to 5 MeV with a precision of about 0.1 per cent. The

whole apparatus is enclosed in a pressure tank and operated at about 10 atmospheres. In this form, the maximum energy is close to 10 MeV and is limited by breakdown between the terminal and its surroundings. However, double the energy can be attained by accelerating negative ions to the positively charged terminal; then, through electron-stripping, positive ions are created which can be accelerated again as they pass from the terminal to ground. Such "tandem" Van de Graaff generators, in two- and three-stage variations, can provide particles in the 10 to 30 MeV range, with high precision.

An electric field may also be produced by a time-varying magnetic field and such a semidirect method of application is used in a *betatron*. The changing magnetic flux in the central core of a pulsed cylindrical electromagnet induces a transverse electric field that accelerates the particles. These travel in a doughnut-shaped vacuum chamber located between the poles of the magnet surrounding the core. The magnetic field between these poles keeps the particles traveling in a circle, but it must be carefully designed to keep the particle orbits within the vacuum chamber during each pulse. Although betatrons can accelerate positively charged particles, they have been used exclusively for electrons. The electrons can be extracted, but they usually bombard an inner target to produce beams of x-rays which can be as intense as 14000 R/min at a distance of 1 meter from the target. Pulsing rates vary from 30 to 60 times per second. The highest energy that has been achieved in a betatron is 300 MeV (University of Illinois), but the majority are designed for the range around 25 MeV. An upper limit to the energy arises from the fact that electrons, when traveling on circular orbits at relativistic velocities, radiate electromagnetic energy. These radiation losses disturb the particles' orbits, and compensation must be provided. This becomes impractical in betatrons above about 300 MeV.

All other types of accelerators use various forms of rf electric fields, at relatively low voltage, which are applied many times in a given direction to the particles and are prevented from influencing them when the rf field is reversed.

One way of applying this principle is used in *linear accelerators* ("linacs") where the particles travel in a straight line down the center of a cylindrical pipe that acts as a waveguide. The waveguide has a rf electromagnetic field pattern whose axial electric-field component provides the accelerating force. To keep the length to a reasonable size, quite high fields (several million volts per meter) are needed with associated high power requirements. This means that the linear accelerator is relatively more costly for a given energy than other accelerators. However, it does have the advantage that a well-focused beam, with small energy spread, can emerge and be readily available for experiments.

The structures of linear accelerators for electrons differ quite markedly from those for protons and heavy ions due to the fact that electrons reach very nearly the speed of light at energies of

* The energy acquired by the particles in an accelerator is expressed in electron volts (eV), the amount of energy gained by any particle bearing a charge equal to that of an electron when it falls through a potential difference of 1 volt. $10^3\text{eV} = 1\text{keV}$; $10^6\text{eV} = 1\text{MeV}$; $10^9\text{eV} = 1\text{GeV}$ (U.S.A.) or 1 GeV (Europe).

a few million electron volts, whereas protons do not reach this velocity until they are accelerated to the billion-electron-volt range. In a *proton linear accelerator*, the waveguide is a resonant cavity where the particles are shielded from a standing rf wave, during the part of the cycle not suitable for acceleration, by a series of "drift tubes," and then are accelerated when traveling across gaps between the drift tubes. As the protons' velocity increases, the drift-tube length increases so that the particles which arrive at a gap at an accelerating phase continue to be accelerated at succeeding gaps. The protons are injected into the linac at energies from 500 to 1000 keV, usually from a Cockcroft-Walton generator. The proton linac is capable of delivering currents of several milliamperes, but usual operation is limited to short pulses (10 to 100 μ sec) with a duty cycle of 1 per cent or less. In principle, there is no limit to the energy, but to date (1965) 70 MeV is the highest achieved, although design studies have been carried out for machines up to 1 BeV. Several proton linacs, of 10 to 50 MeV, are in use as injectors for high-energy synchrotrons. Heavy-ion linacs, similar in design to proton machines, have been built to give energies of about 10 MeV/nucleon.

The waveguide in an *electron linac* is a relatively simple pipe consisting of many small cavities, and the electrons travel in step with the electromagnetic wave. The cavities are coupled by a series of irises (hollow disks) that are constructed to give the proper phase velocity. The whole structure requires extremely close tolerances and strict temperature control. It is usual to use a traveling wave mode of operation although this is not essential. Many electron linacs are in operation up to about 100 MeV, and 1-BeV machines are operating at Stanford University, U.S.A. and at Orsay, France. Average currents up to 1 μ A have been obtained at pulse rates of 60/sec. Under construction, at Stanford, is a 2-mile-long electron linear accelerator for an energy of about 20 BeV in its first stage of operation; then, with the addition of more rf power sources, it will attain about 40 BeV.

Another method of applying relatively small voltages many times to raise the energy of particles is used in the cyclic accelerators. A magnetic field is applied in a direction perpendicular to the particles' plane of motion resulting in a spiraling or circular orbit. Single or multiple rf sources, located on such an orbit, provide increments in energy on each revolution, and the particles continue to circulate until the design energy is attained. In such motion, the magnetic force must balance the centrifugal force, i.e.,

$$mv^2 = evB \quad (1)$$

where e is the charge, m the mass, and v the velocity of the particle, r is the radius of the path, and B is the magnetic flux density.

In a *cyclotron*, the magnetic field is almost uniform over the region between two cylindrical

poles, and it is constant in time. Positive ions originate in an ion source located near the center of the gap and are accelerated across the gaps between two D-shaped electrodes which each cover almost half the area of the magnet poles. Between the "dees" a constant frequency rf field is applied. From Eq. (1), it can be seen that the angular velocity, v/r , of the particle is given by eB/m , that is, it is independent of radius. Thus, the particles always take the same time to cover the distance between the accelerating gaps as they travel on an orbit of half-circles of ever-increasing radius; they arrive at each gap always at the proper phase to continue to be accelerated. The cyclotron is most useful for accelerating protons and deuterons in the energy range between about 5 and 20 MeV, but heavy ions such as multiply ionized carbon, nitrogen, etc., can be accelerated to about 100 MeV. Continuous currents of about 1 mA can be obtained for use with internal targets, or about one-tenth of this can be extracted for external use by means of a deflecting field situated at the outer radius. There is an upper limit in the energy obtainable in a cyclotron because the relativistic increase in mass of the particles as their energy increases causes them to reach the accelerating gap progressively later and to fall out of resonance with the accelerating field. This limit is about 30 MeV for protons and 40 MeV for deuterons.

In the *synchrocyclotron*, this basic limitation is removed by varying the frequency of the accelerating field, decreasing it to keep in step with the decreasing frequency of revolution of the circulating particles. However, even with a wide variation in frequency, the particles can remain at or near the correct acceleration phase. A particle which receives too much energy will travel on a path with a larger radius than the "equilibrium" particle that receives just the right amount, thus the higher-energy one will return to the accelerating gap somewhat later in phase. Phase stability will result if the "equilibrium" phase is on the falling side of the rf wave, i.e., between 90 and 180°. Synchrocyclotrons have been constructed to accelerate positive ions (chiefly protons) to energies from about 100 to 700 MeV, the latter being about the practical maximum due to the size of the magnet. This is similar to that of the cyclotron, the almost constant magnetic field being produced between cylindrical poles. For a 700-MeV design, such a magnet weighs about 7000 tons and the frequency must drop to less than 60 per cent of its initial value. The tuning for the rf systems is provided either by a rotating condenser or a vibrating-reed type of variable capacitor. With such frequency modulation over the accelerating period, it is no longer possible to have continuous acceleration as in the conventional cyclotron. The output of a synchrocyclotron is pulsed and time-average internal currents are only about 1 μ A. Cycling rates vary from 30 to 100 cps.

The *microtron* is a variation of the cyclotron that is used to accelerate electrons. Again, the magnetic field is constant and the accelerating rf field has a constant frequency. The accelerating gap is

now situated at one edge of the magnet, and the circular orbits of the particles are tangent at this point. The amount of energy given to the particles is adjusted so that the time of revolution in each orbit is one or more rf periods longer than in the previous orbit. Thus, the particles return to the accelerating region at the same phase on each turn. Several microtrons have been built for energies from a few million electron volts to almost 30 MeV, with currents of about $1/2 \mu\text{A}$.

The highest energies, for both electrons and protons, have been reached in the *synchrotron*. In this type of accelerator, the particles are kept moving on an orbit that is almost circular, and of a fixed radius, between the poles of a magnet that is annular in shape. For the higher energies, such a magnet is much less costly than the cylindrical synchrocyclotron magnet. From Eq. (1), above, it can be seen that to keep the particles' orbit with a constant radius, the magnetic field must be increased during the period of acceleration from a low value that would correspond to the injection energy to a final value corresponding to the maximum energy. The radius of the accelerator is determined by the value of this maximum energy and the maximum value obtainable for the magnetic field (usually 10 to 15 kilogauss). Accelerating fields are provided by one or more rf stations located at points on the magnet ring, and the frequency must increase exactly in step with the increasing velocity of the particles.

In all cyclic accelerators, focusing forces to keep the particles within the limits of the vacuum chamber are provided by shaping the magnet poles to give a field that varies with radius. Synchrotrons fall into two general categories: (1) weak-focusing, constant-gradient (CG) synchrotrons where a small radial negative gradient has a constant value in azimuth and (2) strong-focusing alternating-gradient (AG) synchrotrons where large radial gradients alternate in sign with azimuth. From the equations of motion of particles in an azimuthally uniform field, the radial field index n [$n = -(dB/dr)(r/B)$] must have values between 0 and 1 for stability in both radial and vertical directions. Stronger gradients provide strong focusing in one plane but result in defocusing in the other plane. Alternating-gradient systems depend upon the principle that a focusing combination can result from an alternation of focusing and defocusing sections (as in optical lenses). In almost all synchrotrons, of either type, the ring magnet is broken into several (or many) sections separated by field-free regions that are used for injection, for rf accelerating stations, for beam observation and control mechanisms and for targeting.

Electron synchrotrons range in energy from a few tens of million electron volts up to somewhat over 1 BeV in the constant-gradient type and up to 6 BeV in the alternating-gradient variety. As in the betatron, the practical limit in energy is set by radiation losses which amount to several million electron volts per revolution in the 6-BeV machines. Since the velocity of electrons is so close to that of light, above a few million electron

volts, the accelerating frequency can be constant. To reach this energy, the smaller electron synchrotrons include some fluxbars inside the orbit to provide betatron-type acceleration and the larger ones use another accelerator as an injector such as a Van de Graaff generator or electron linac. The output from most electron synchrotrons is in the form of a continuous spectrum of x-rays, obtained when the beam strikes an internal target, or the beam can be extracted. These accelerators are usually pulsed at rates from 20 to 60 cps.

Proton synchrotrons have been built only for energies of 1 BeV and higher, since the synchrocyclotron can accelerate positive particles up to several hundred million electron volts. The constant-gradient type requires large magnetic apertures, vacuum chambers several feet wide, and massive magnets with cross sections of many square feet. It is probable that the highest energy CG proton synchrotron that will be built is the 12.5-BeV machine now operating at the Argonne National Laboratory. In the alternating-gradient synchrotrons, the stronger forces keep the protons within a cross-sectional area of a few square inches, resulting in a simple vacuum chamber and magnets about 3 feet by 3 feet, or less, in cross section. However, the stronger forces necessitate extreme precision in location and stability of the whole structure and stricter controls. The highest energy achieved to date (1964) is the 33-BeV AG proton synchrotron (half-mile circumference) at the Brookhaven National Laboratory, but a 70-BeV accelerator is under construction near Moscow, U.S.S.R. There seems to be no basic limitation to the energy achievable in AG proton synchrotrons, except cost, and design studies have been made for such machines up to 1000 BeV.

Because protons do not reach velocities close to that of light until they reach BeV energies, the accelerating systems must have a very wide frequency range that increases rapidly from the injection value to many times this value, and at the same time it must provide the correct value corresponding to the particles' energy to very high precision. In the early CG proton synchrotrons, the rf frequency is that of the frequency of revolution of the particles, but in newer accelerators, both CG and AG, a higher harmonic is used. The location and behavior of the beam is observed by placing, adjacent to the beam, "pickup electrodes" upon which charge is induced. Signals fed back from these electrodes can be used to provide corrections to the accelerating frequency program. Such control is essential in AG accelerators and is also used in some CG machines.

Because of the difficulty in providing sufficiently accurate magnetic fields at low values, other types of accelerators are used to raise the protons' energy to several MeV before injection. In the smaller CG proton synchrotrons, the injectors are usually Van de Graaff electrostatic generators of 3 to 4 MeV, but the larger synchrotrons, both CG and AG, use proton linear accelerators of 10 to 50 MeV. The time taken to complete acceleration

to the full energy is of the order of 1 second, and repetition rates vary from 10 to 20 pulses/min. Typical intensities in proton synchrotrons are 10^{11} to 10^{12} protons/pulse, and space-charge effects set an upper limit at intensities about an order of magnitude, or so, higher than this. Recently various types of instabilities have been observed at intensities below these theoretical values and the actual limits are still under investigation.

With alternating-gradient focusing, particle orbits are quite closely spaced even for large differences in momentum, and this has led to the development of Fixed-Field Alternating-Gradient (FFAG) accelerators. In these, the magnetic field can remain constant in time, but the annular magnet must have considerably greater radial width than the synchrotron's in order to contain all orbits from the injection energy to the desired top value. In the radial-sector FFAG design, the magnet is broken into wedge-shaped sectors where the magnetic field increases strongly with radius but is reversed in sign in alternate sectors. In the spiral-ridge design, the magnetic field again increases rapidly with radius and remains constant in sign, and spiraling ridges on the magnet poles give regions of stronger field. These ridges provide alternations in focusing forces as the particles cross them during their revolutions. Although these accelerators are much more costly than the pulsed synchrotrons, they can provide much higher intensities. This can be accomplished by accelerating particles to a high energy and "stacking" many groups of them at their high-energy orbit. No FFAG synchrotron has been constructed as yet for high energies although models have demonstrated their feasibility.

A variation of this principle has been used in the construction of AVF (Azimuthally Varying Field), or Sector-Focused, Cyclotrons in which the cylindrical magnets have four, or more, slightly spiraling sectors that give alternations in gap height and thus in field. The strong focusing forces and an increase in magnetic field with radius allow the particles to stay in resonance as their mass increases. Several AVF cyclotrons have been constructed for positive ions, with proton energies up to 80 MeV, and designs have been proposed up to almost 1 BeV.

As the construction of accelerators has progressed to higher and higher energies and the particles become more relativistic, one source of concern has been the limitation on the actual fraction of energy available in the center-of-mass system for the production of new particles and interesting reactions. For example, when protons with kinetic energy of 30 BeV strike a fixed target, only 5.86 BeV is available for particle production. This has led to a consideration of providing beams of equal and opposite momenta which can intersect at one or more locations and so provide their total kinetic energy in the center-of-mass system. To obtain a significant number of interactions, high intensities are required. Designs have been studied for intersecting magnetic storage rings where many pulses from existing high-energy proton synchrotrons could be collected for such

use, but none have been built as yet (1964). However, storage rings for electrons of around 500-MeV energy have been constructed at Stanford University, U.S.A., and in Europe, and investigations with these will provide information concerning the usefulness of such devices.

M. HILDRED BLEWETT

References

- Livingston, M. S., and Blewett, J. P., "Particle Accelerators", New York, McGraw-Hill Book Co., 1962.
 Livingood, J. J., "Principles of Cyclic Particle Accelerators", Princeton, N.J., D. Van Nostrand, 1961.
 Flügge, S., Ed., "Handbuch der Physik," Vol. 44, Berlin, Springer, 1959.

Cross-references: ACCELERATORS, LINEAR; ACCELERATORS, VAN DE GRAAFF; BETATRON; CYCLOTRON; SYNCHROTRONS.

ACCELERATORS, VAN DE GRAAFF

The electrostatic particle accelerator originated by American physicist Robert Jemison Van de Graaff is widely used for nuclear structure research. These constant-potential accelerators make use of the electrostatic belt generator invented by Van de Graaff about 1930. They belong to the *direct* accelerator family in which the high voltage power is applied directly across the terminals of a highly evacuated multi-electrode tube. Electrified atoms or electrons from a source within the high-voltage terminal gain velocity and energy as they move along the tube axis to ground under the action of the applied electric field. As each particle emerges from the accelerator, it is moving with a kinetic energy equal to qV where q is the particle charge and V the generator voltage.

While a Rhodes Scholar at Oxford during 1927 and 1928, Van de Graaff selected the electrostatic approach to fulfill the need, much emphasized by Rutherford, for more copious sources of atomic particles comparable in energy to those spontaneously emitted from naturally radioactive materials (see ELECTROSTATICS). Subsequently, at Princeton University, Van de Graaff produced over one million volts between the spherical terminals of two small electrostatic belt generators of a new and surprisingly simple design; in 1931, he described the electrostatic belt generator principles, and their suitability for the bombardment of atomic nuclei, before the American Physical Society. The method was first applied to nuclear investigations at the Carnegie Institution of Washington in 1932. The early machines, insulated in atmospheric air, produced streams of light positive ions such as protons and deuterons homogeneous in energy and with smooth control over the voltage range of the machine. General acceptance of the Van de Graaff accelerator as the precision instrument for experimental nuclear research followed rapidly, and its further development for this purpose has been continuous since

that time. Greater compactness and higher voltage were attained by insulating the belt generator and tube with compressed gas; greater beam intensity came through improved ion source and acceleration tube technology. About 300 such accelerators were in use by 1960, producing particles and radiation with energies from 400 keV to 10 MeV. At that time, Van de Graaff accelerators for nuclear science incorporated the "tandem acceleration" principle described below. It opened the way to far higher particle energies by applying the tandem principles to multiply charged heavy atoms.

Van de Graaff accelerators can accelerate any electrified particle, including any of the 92 elements, electrons, and clumps of matter simulating micrometeorites. In addition to use in experimental nuclear physics with high-energy positive ions, Van de Graaff electron accelerators designed for voltages in the 1 to 5 MeV range are used to produce megavolt x-rays for the treatment of malignant disease and for the radiographic inspection of heavy opaque structures such as metal forgings, weldments, and rocket engines. Streams of electrons from such accelerators are also used for radiobiological and radiochemical research, and for the treatment of skin malignancies. Radiation processing studies for such

purposes as the sterilization of surgical materials, the cross-linking of polyethylene and other plastics, the deinfestation of grains, and increased shelf life of foods have often made use of Van de Graaff accelerators.

Van de Graaff Generator Operating Principles. Although a variety of electrostatic machines had been developed since the first frictionally excited generator of Otto von Guericke in the middle of the seventeenth century, all have been superseded by the Van de Graaff generator¹ because of its greater voltage capability and comparative simplicity. The essential components of the generator, outlined in Fig. 1, include a well-rounded metal terminal supported by an insulating column and an endless insulating belt system which physically conveys electric charge from ground to the high voltage terminal.

Electric charge of the desired polarity is deposited on the moving belt surface by corona from a row of metal points at a controllable voltage with respect to the lower pulley toward which they are directed. In addition to overcoming friction and windage, the motor-driven belt does work in carrying this charge from ground to the terminal potential. Transfer of the charge from belt to terminal is accomplished by again presenting a row of points toward the electrified belt. This

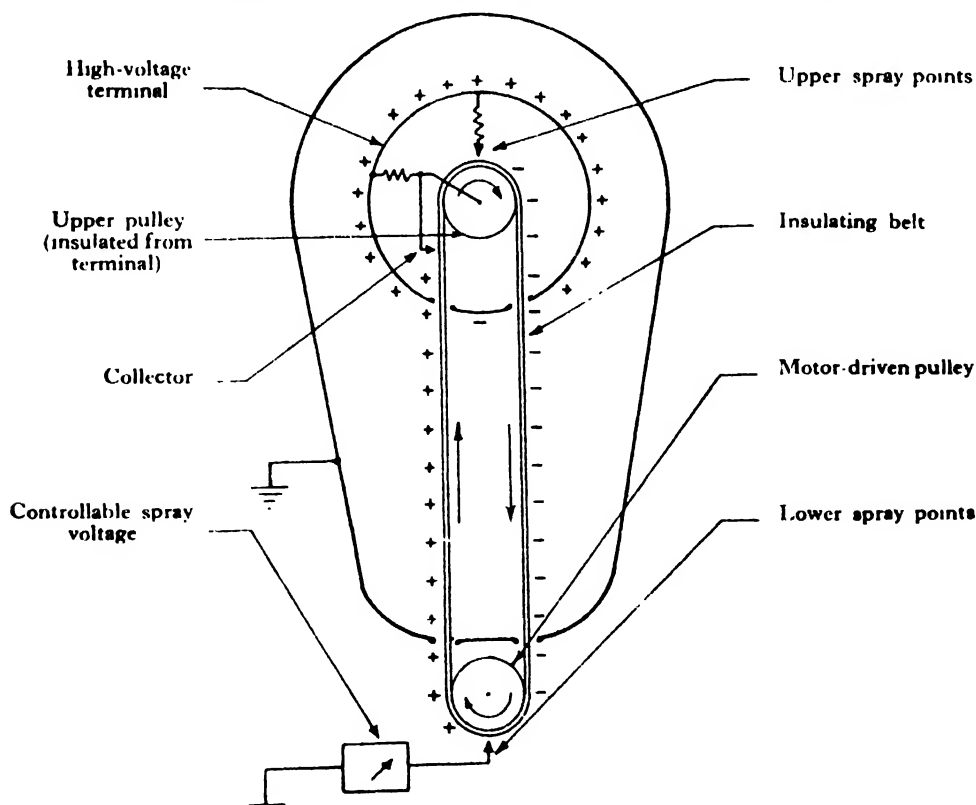


FIG. 1. Diagram of Van de Graaff electrostatic belt generator. Reproduced by permission of The Institute of Physics and The Physical Society from the article by R. J. Van de Graaff, J. G. Trump, and W. A. Buechner, "Reports on Progress in Physics," 11, p. 1, 1948.

time the electric field of the surface-bound charge produces the gaseous ionization needed for conduction across the point-to-belt gap. Van de Graaff pointed out that these ionized charge-transfer processes remain independent of the terminal voltage if they are located in the field-free space within the hollow terminal or below the ground plane. The current of such an electrostatic generator is limited by the maximum charge density which can be insulated in the gaseous medium surrounding the belt and the total area per second of charge-laden surface entering or leaving the terminal. To increase the current capability of the system, the return run of belt may be charged within the terminal in a similar manner but with the opposite polarity.

The potential, V , of the high voltage terminal of a Van de Graaff generator is determined by the amount and polarity of the accumulated charge on its insulated terminal. At any instant $V = Q/C$ where Q is the net positive or negative charge on the terminal and C is the capacitance of the terminal system to ground. Although the Van de Graaff generator is inherently a constant-current machine, it can be maintained steadily at the desired voltage by balancing the current arriving at the terminal against the total current delivered to the load. The load usually includes the particle current through the accelerating tube, the current through resistors which divide the terminal voltage uniformly along the supporting column, and any corona from the terminal itself arising from the high electric field at its surface. By adjusting either the belt current or the load current, the terminal voltage may be maintained at any desired value up to the maximum which can be insulated. This maximum voltage depends only on the physical size and geometry of the terminal and on the electrical strength of dielectric medium surrounding it. An isolated metallic sphere would be the ideal terminal, but modifications are necessitated by the supporting column, belt and tube.

The pair of generators built by Van de Graaff at Princeton in 1930 each had an aluminum spherical terminal 2 feet in diameter supported in air by a slender glass rod 7 feet long. A silk ribbon was employed as the insulating charge conveyor. The voltage insulated in atmospheric air between these two generators, one accumulating positive and the other negative charge, was more than twice any previously attained constant voltage.

About 5.5 million volts were insulated in air between two larger generators constructed by Van de Graaff in the early 1930's for nuclear research. This voltage required spherical terminals 15 feet in diameter supported on insulating tubular columns 25 feet high. This historic equipment, shown in an early sparking demonstration in Fig. 2, was used in a modified form for precision nuclear research at Massachusetts Institute of Technology for nearly 20 years. It is now installed at the Boston Museum of Science for demonstrations of the principles and phenomena of electrical science.

The need for still higher constant voltages for

nuclear investigations, and the desire for more compact apparatus, led to the use of high-pressure gases for the insulation of electrostatic accelerators. Today nearly all Van de Graaff accelerators operating at potentials in excess of one-half million volts are within a steel pressure tank and insulated in gases compressed to 10 to 25 atmospheres. Electronegative gases such as sulfur hexafluoride (SF_6) and "Freon" (CCl_2F_2) are now increasingly used instead of mixtures of nitrogen and CO_2 , since they insulate approximately the same voltages at one-third gas pressure.

Acceleration System. The evacuated acceleration tube, the source of positive ions or electrons, and the target to which the energized particles are directed, constitute the particle accelerating system of the Van de Graaff accelerator. The insulating length of the evacuated acceleration tube is divided into many sections by metal disk-like electrodes, each with an axial opening for the passage of the particle beam. Each disk is mounted between annular rings of glass or porcelain to form a slender vacuum-tight accelerating column. The tube electrodes take their potential from the metallic members in the generator column along which the terminal voltage is divided by resistors. The charged particles, acted upon by the electric field between these electrodes, are progressively accelerated and focused as they move through the electric fields between the electrodes. At the remote end, the beam emerges as a collimated and directed stream of energetic particles.

Tandem Acceleration and Multiply Charged Ions. Van de Graaff accelerators for nuclear science now reach higher particle energies with a given terminal voltage by switching the polarity of the accelerated particles. In the two-stage tandem diagramed in Fig. 3, negatively charged ions are produced at ground and then accelerated toward a high-voltage positive terminal. Within this terminal, the swiftly moving negative ions are stripped of electrons by passing through a thin gaseous region. The resultant positive ions continue through the tube under the second accelerating action of the positive terminal. A singly charged particle, such as a proton, thus arrives at the ground end of the system with an energy of $2qV$.

At sufficiently high energy, atoms of higher atomic number may be stripped of several or even of all their satellite electrons. An ion which lacks N electric charges during the second acceleration stage gains a total energy of $(N + 1)V$ in a two-stage tandem accelerator. Three-stage acceleration is secured by adding an additional in-line two-stage accelerator with a central negative terminal and using it to produce one stage of negative ion acceleration for injection into the second tandem. Although the light elements, hydrogen and helium, were almost exclusively used as atomic projectiles in nuclear structure physics until 1960, interest in heavier nuclei developed rapidly as higher energies became possible. It is estimated that, by applying tandem acceleration principles, a two-stage



FIG. 2. 5.5-million volt Van de Graaff generator in sparking demonstration.

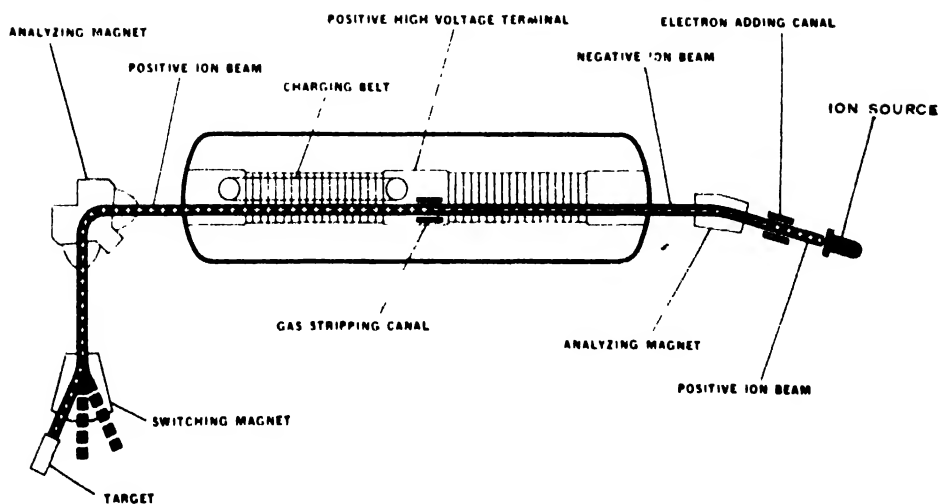


FIG. 3. Diagram of two-stage tandem Van de Graaff accelerator. Reproduced from the article by R. J. Van de Graaff in "Nuclear Instruments and Methods," 8, p. 195-202, 1960, by permission of the North-Holland Publishing Co.

Van de Graaff accelerator with a 15 MeV positive terminal can produce a beam of uranium ions with energies up to 400 MeV. In large part because of the more complete electron stripping attained at higher energies, three-stage acceleration could produce uranium ions with energies over 1000 MeV.

Medical Applications of Van de Graaff Accelerators. Since x-rays are the form of electromagnetic energy, similar to light, produced by the sudden stopping of high-energy electrons, Van de Graaff accelerators are often used as x-ray sources for the treatment of malignant disease and for radiography. In this application, the high-voltage terminal is operated at negative polarity, the electrons are emitted from a tungsten source at the terminal end of the acceleration tube, and they are suddenly stopped after traversing the length of the tube by striking a water-cooled metal target, usually of tungsten or gold.

A 2 million volt x-ray generator of this type, in which a gold target is bombarded with 300 μ A of electrons, yields an x-ray intensity of 100 r/min measured 1 meter from the target in the electron direction. The quality of this radiation is closely similar in its physical properties to that of the gamma rays from radium or from the radioactive isotope cobalt 60. To equal this x-ray intensity would require over 4000 curies of cobalt 60 or 6000 grams of radium. This Van de Graaff accelerator for therapy is housed in a steel tank 3 feet in diameter and 6 feet long and is insulated by a mixture of nitrogen and CO₂ at 300 psi.

JOHN G. TRUMP

References

- Van de Graaff, R. J., Trump, J. G., and Buechner, W. W., "Electrostatic Generators for the Acceleration of Charged Particles," *Rept. Progr. Phys.*, **11**, 1 (1948).
 Van de Graaff, R. J., "Tandem Electrostatic Accelerators," *Nuclear Instr. and Methods*, **8**, 195-202 (1960).
 Wittkower, A. B., Rose, P. H., Bastide, R. P., and Brooks, N. B., "Injection of Intense Neutral Beams into a Tandem Accelerator," *Rev. Sci. Instr.* **35**, 1-11 (January 1964).
 Wright, K. A., Proimos, B. S., and Trump, John G., "Physical Aspects of Two Million Volt X-ray Therapy," *Surg. Clin. North Am.*, **39**, 1-12 (June 1959).
 Livingston, M. Stanley, and Blewett, J. P., "Particle Accelerators," Ch. 3, New York, McGraw-Hill Book Co., 1962.

Cross-references: ACCELERATORS, LINEAR; ACCELERATORS, PARTICLE; CYCLOTRON; STATIC ELECTRICITY; SYNCHROTRON.

ACOUSTICS

Physical acoustics deals with the properties and behavior of longitudinal waves of "infinitesimal" amplitude in solid, liquid, or gaseous media. These waves are propagated at the velocity of

sound, or phase velocity, which is independent of frequency in a non-dissipating free medium. In such a case, the shape of a complex wave remains unchanged during its propagation, although its amplitude may change. When the velocity of sound, or phase velocity, becomes dependent on frequency, the shape of a complex wave changes during propagation and dispersion is said to occur. In such cases groups of waves comprising a limited range of frequencies travel at a velocity called the group velocity, different from the phase velocity. It is the group velocity which carries the energy of such complex waves.

Acoustic waves are dispersive (1) in a *free* medium in which viscosity, heat conduction, and molecular, thermal, or chemical relaxation cause an increase in phase velocity with frequency, (2) in a *confined* medium in a capillary tube in which viscosity causes a decrease in phase velocity with frequency, (3) in a *confined* medium in *non-dissipative* tubes of increasing cross section, where the rate of change of cross-sectional area differs from the conical (i.e., different from proportionality to the square of the distance along the tube)—examples of such tubes being the exponential and catenoidal horns, in which the phase velocity increases with decreasing frequency, (4) in non-dissipative cylindrical tubes with *flexible* walls, and (5) in waves of *finite* amplitude, where the higher-frequency components have a higher phase velocity than the lower-frequency components, a transfer of energy occurring from the lower-frequency components to the higher-frequency components.

In physical acoustics, waves are reflected, refracted, diffracted, and absorbed. They exhibit all the properties of wave motion, such as reinforcement and destructive interference. They are accompanied by pressure and particle-velocity fluctuations detectable by the ear or by instruments capable of measuring the frequency instantaneous values, and mean intensity of these fluctuations.

Geometrical acoustics is a special case of physical acoustics in which diffraction and interference are disregarded. Energies of direct and reflected waves are considered to add irrespective of relative phase, a condition applicable to incoherent (i.e., uncorrelated) waves.

Electroacoustics deals with the manner and means of energy transfer between electromechanical devices and the medium propagating sound waves. It is concerned with microphones of various types (dynamic, electrostatic, piezoelectric, magnetostrictive), loudspeakers of equivalent types (to which Maxwell's reciprocity theorem applies), and amplifiers for the faithful transfer of signals received by a microphone to those reproduced by a loudspeaker, tape recorder, or oscilloscope.

Architectural acoustics deals with the problems of distribution of beneficial sounds within buildings and with the exclusion or reduction of undesirable sounds. Here it is shown that mass and limpness of barriers such as partitions are most significant in providing high sound transmission

loss. It is also shown that sound transmission loss tests on relatively small partitions (of the order of 66×80 inches) often give significantly higher transmission loss values than the full size partitions (of the order of 100×180 inches) usually used in practice. This points up the essential need to rely on full-scale tests, rather than on smaller-scale tests, for sound insulation data. Moreover, the application of sound absorbing materials, such as acoustical tiles or fibrous sound-absorbing blankets, to walls or ceilings, have a minor effect on transmission loss through partitions, vertical or horizontal. This applies equally well to such remedial measures as blowing rock wool between studs in walls or between joists in floors or ceilings. The only significant improvement in sound transmission loss results from massive structures or from discontinuous constructions such as floating floors, heavy flexibly suspended ceilings, staggered studs or multiple leaves in walls.

In auditoriums, reflective ceilings and reflective walls, combined with convex irregularities of random design, provide for reinforcement and diffuseness of sound found so beneficial for speech and music. Reflecting surfaces, giving short time-delay reflections (about 20 msec or less), are particularly desirable in concert halls. Delays of 65 msec or more may result in echoes and speech unintelligibility.

Psychological acoustics deals with the emotional and mental reactions of persons and animals to various sounds. Here questions arise as to which sounds are acceptable to most people under various living conditions and which are not. For this purpose, various noise criteria have been developed, related to the so-called speech interference level (SIL). By definition, SIL is the average of the sound pressure levels in decibels (L_p , see below) in the three octave frequency bands 600 to 1200 cps, 1200 to 2400 cps, and 2400 to 4800-cps. The loudness level in phons (L_N , see below) of a broad-band noise (no outstanding pure tones) should be not over 22 phons (at the most, not over 30 phons) greater than the SIL, in decibels, of the background noise. Two noise control criteria, so-called NC and NCA, are designed to fulfil these conditions for various sound fields ranging from radio broadcasting studios, through bedrooms, offices, restaurants, sports arenas and factories.

Physiological acoustics deals with hearing and its impairment, the voice mechanism, and the physical effects in general of sounds on living bodies.

The frequency range of sound is divided into three somewhat overlapping regions, namely, an *audio-frequency* band ranging from approximately 20 to 20000 cps flanked by an *infrasonic* region below 30 cps and an *ultrasonic* region above 15000 cps. Human ears do not respond in general to frequencies outside the audio band, although small animals such as cats and bats do hear in the lower ultrasonic region. At one time called *supersonics*, the term *ultrasonics* is now accepted to distinguish this area of high-frequency

sound propagation from the cases of supersonic aircraft, supersonic fluid flow, and shock waves in fluids, which have to do with speeds higher than the speeds of sound.

The *strength* of a sound field is measured by its mean square pressure expressed as sound pressure level (L_p) in decibels. Decibels are logarithmic units defining the range of sound pressure levels (L_p) between the minimum audible value at 1000 cps (4db*—the threshold of hearing) for the average pair of good young (high school age) ears and the maximum audible value of L_p at which effects other than hearing (such as tickling in the ears—the threshold of *feeling*) begin to appear. This upper limit shows up at about 120 db at 1000 cps.

Higher values of L_p (e.g., 130 db) begin to cause pain in the average ear, and values of 160 db may well cause instantaneous physical damage (perforation) to the tympanic membrane. The minimum audible sound pressure, p_0 , at 1000 cps is internationally accepted as 0.0002 microbars rms (i.e., 0.0002 dynes/cm², rms), and the sound pressure level at any other rms value of sound pressure, p , irrespective of frequency, is given by $L_p = 20 \log_{10} (p/p_0)$ db. The bel (seldom used) is simply equal to 10 decibels. The bel appears first used in connection with power loss in telephone lines and is named in honor of Alexander Graham Bell.

Other reference pressures, p_0 , may be used in special applications, instead of 0.0002 microbar, so it is essential to specify the reference pressure when quoting values of L_p .

The *loudness* of a sound field is judged by the ear in the audio frequency range. Loudness judgments by groups of observers have established a *loudness level* scale. The loudness level (L_N) in phons is arbitrarily taken equal to the sound pressure level L_p in db at the reference frequency of 1000 cps over the range from the threshold of hearing to the threshold of feeling. Jury judgment of equality in loudness between test tones at different frequencies (f) and 1000-cps reference tones of known sound pressure level (L_p) have established *equal loudness* contours (contours of constant L_N) in the L_p - f plane.

These contours show in general a marked decrease in ear sensitivity to sounds at frequencies below about 200 cps, and this decrease is much more pronounced in the lower loudness levels. For example, at 50 cps the 4-phon contour has an L_p of about 43 db, the 80-phon contour about 93 db. At higher frequencies, the ear shows some 8-db increase in sensitivity in the region around 3500 cps, then a loss in sensitivity beyond about 6000 cps. These characteristics of hearing are significant in the design of lecture and music halls, noise-control devices, and high-fidelity audio equipment.

Also based on jury judgments, a scale of *loudness*, N , (in sones) has been established for sounds (for pure tones and for broad band noise). On

* 0 db at 1000 cps is defined in older work as the threshold of hearing.

this scale, a given percentage change in some value denotes an equal percentage change in the subjective loudness of the sound. The scale provides single numbers for judging the relative loudnesses of different acoustical environments, for evaluating the percentage reduction in noise due to various noise control measures, and for setting limits on permissible noise in factories, from motor vehicles, etc.

Loudness N is related to loudness level L_N in the range 40 to 100 phons by the equation $\log_{10} N = 0.03 L_N - 1.2$. A loudness of 1 sone corresponds to a loudness level of 40 phons and is typical of the low-level background noise in a quiet home.

Various methods are available for estimating loudness of complex sounds from their sound pressure levels in octave, half-octave, or third-octave bands. For traffic noises, readings on a standard sound level meter using the A-scale (which incorporates a frequency-weighting network approximating the variation of ear sensitivity with frequency to tones of 40-db sound pressure level) appear to correlate reasonably well with jury judgments of vehicle loudness.

The *noisiness* of a broad-band noise is more related to the annoyance it causes than to its loudness. Thus, corresponding to the scale of sones created to measure loudness, a scale of *noys* has been developed as a measure of the noisiness of jet aircraft noise in particular. Noys give more importance to the high-frequency bands of noise and less importance to the low-frequency bands than do sones. Also, corresponding to the scale of loudness levels in phons, there has been established a scale of *perceived noise levels* in PN db. Rules have been established for converting sound pressure level measured in octave bands, half-octave bands and third-octave bands into noys and then into PN db. Although originally developed as a means for the assessment of the "noisiness" of jet aircraft flying over inhabited communities, the concept of noisiness is being applied to traffic and other broad band noises.

W. W. SOROKA

References

Books

- Beranek, Leo L., "Acoustics," New York, McGraw-Hill Book Co., 1954.
 Harris, Cyril M., "Handbook of Noise Control," New York, McGraw-Hill Book Co., 1957.
 Beranek, Leo L., "Noise Reduction," New York, McGraw-Hill Book Co., 1960.
 Officer, C. B., "Introduction to the Theory of Sound Transmission," New York, McGraw-Hill Book Co., 1958.
 Ewing, W. Maurice, Jardetzky, Wenceslas S., and Press, Frank, "Elastic Waves in Layered Media," New York, McGraw-Hill Book Co., 1957.
 Morse, Philip M., "Vibration and Sound," Second edition, New York, McGraw-Hill Book Co., 1948.
 Kinsler, Lawrence E., and Frey, Austin R., "Fundamentals of Acoustics," Second edition, New York, John Wiley & Sons, 1962.

Periodicals

- Journal of the Acoustical Society of America* (1929-).
Acustica (1951-).
Proceedings of the International Congresses on Acoustics.
Journal of the Audio Engineering Society (1953-).
Journal of Sound and Vibration (1964-).

Cross-references: ARCHITECTURAL ACOUSTICS; ELECTROACOUSTICS; HEARING; MUSICAL SOUND; NOISE; ACOUSTICAL; PHONONS; PHYSICAL ACOUSTICS; REPRODUCTION OF SOUND; RESONANCE; SONAR; ULTRASONICS; VIBRATION; WAVE MOTION.

ADSORPTION AND ABSORPTION

When a porous solid such as charcoal is exposed, in a closed space, to a gas such as ammonia, the pressure of the gas diminishes and the weight of the solid increases; this is an example of the adsorption of a gas by a solid. It is termed physical adsorption because the forces bringing it about are the "van der Waals" forces of attraction which act between the molecules of the gas and the atoms or ions comprising the solid. It is now known that all solids, whether porous or nonporous, will adsorb all gases physically, whereas the phenomenon of *chemisorption* is specific in nature. Thus hydrogen is chemisorbed by transition metals such as nickel or iron but not by oxides such as alumina.

In physical adsorption the amount, say w grams, of gas or vapor taken up per gram of solid depends greatly on the nature of the gas and of the solid, as well as on the pressure p and the temperature T . In mathematical form, $w = f(p, T, \text{gas}, \text{solid})$. For a given gas and solid at a fixed temperature, w depends only on pressure, and the relationship between w and p (i.e., $w = f(p)_{T, \text{gas}, \text{solid}}$) is called the *adsorption isotherm*. Data of adsorption are usually quoted in the form of the adsorption isotherm of the gas or vapor under consideration (the "adsorbate") on the given solid (the "adsorbent"). For vapors it is more appropriate to express the isotherm in terms of relative pressure so that $w = f_0(p/p_0)_{T, \text{vapor}, \text{solid}}$, where p_0 is the saturated vapor pressure of the adsorbate at the temperature of the experiment.

With the great majority of solids, the isotherm at the low-pressure end is concave towards the pressure axis. Its further course depends on the nature of the solid: if the solid is nonporous (e.g., a powder) the isotherm reaches a point of inflection at a relative pressure in the region 0.1 to 0.3 and thereafter turns upwards (Type II isotherm); if it is porous, with pores having radii between tens and hundreds of angstroms in diameter, the form of the isotherm is similar, except that at pressures near saturation it bends over and becomes almost horizontal; it also shows a *hysteresis loop* in the middle range of pressures, the desorption branch lying above the adsorption branch (Type IV isotherm). If the solid contains an extensive series of pores of molecular width ($\sim 10\text{\AA}$), the isotherm shows no point of inflection

but continuously diminishes in slope and finally becomes nearly horizontal (Type I isotherm).

Measurements on solids having a known surface area have shown that at low pressures the adsorbed layer is only one molecule thick and as pressure increases the molecules in this "monolayer" become more and more crowded, till at or near the point of inflection in the isotherm the monolayer is completely full up. Further increase in pressure then leads to the formation of a multimolecular layer—a "multilayer"—of gradually increasing thickness.

Adsorption is an exothermic process so that by elementary thermodynamic principles the amount adsorbed by a given solid at a given pressure must diminish as temperature increases. The differential heat of adsorption \bar{q} [i.e., the limit of the ratio $\delta q/(\delta w/M)$ where δq is the heat evolved when the adsorption increases by $\delta w/M$, M being the molecular weight of adsorbate] is rather larger than the latent heat of condensation L of the adsorbate. The value of \bar{q} depends somewhat on the amount adsorbed; it is relatively high for small adsorptions—perhaps about $2L$ —then diminishes as w increases, till in the multilayer region it scarcely exceeds the latent heat. There is usually a low maximum ($\sim 1.3L$ or so) in the region where the monolayer is just complete. The values quoted are only by way of example, and minor, generally unpredictable, differences are found between one system and another. The falling branch of the \bar{q} vs w curve is usually ascribed to heterogeneity of the surface (adsorption occurring first on the more "active" parts of the surface), and the branch rising to the maximum is probably produced by mutual attraction of the molecules within the adsorbed film.

Since physical adsorption results from van der Waals forces, the greater the condensability of the gas or vapor as measured by its boiling point or its critical temperature, the greater is the amount of gas or vapor adsorbed at a given pressure. Thus at room temperature and atmospheric pressure, the "permanent gases" such as hydrogen or nitrogen are only slightly adsorbed even on a good adsorbent such as charcoal, while carbon dioxide is more adsorbed, and benzene and carbon tetrachloride are strongly adsorbed. At very low temperatures the adsorption is correspondingly greater, so that nitrogen at its boiling point of -195°C has an adsorption, on a given solid, comparable with that of benzene at 25°C . For the adsorption to be readily measurable, however, the solid needs to have a relatively large area—a completed monolayer of nitrogen, 1 square meter in extent, weighs only 0.3 mg, for example—so that adsorption phenomena may escape notice unless the solid is "highly disperse," i.e., has an area exceeding several square meters per gram.

Numerous attempts have been made to interpret the detailed course of the different types of isotherm theoretically; but only limited success has been achieved because the models used are necessarily oversimplified and can rarely correspond in detail to the complex systems encoun-

tered in practice. However, it is generally agreed that at or near the point of inflection, the monolayer is complete; thus by assuming a value for the cross-sectional area of an adsorbed molecule, it is possible to estimate from a Type II or a Type IV isotherm the surface area of the solid (the "Brunauer-Emmett-Teller" method). Further, from the course of the desorption loop of a Type IV isotherm, by assuming the adsorbate to be condensed as a liquid in pores which are regarded as cylinders, and by applying the Kelvin equation, one can calculate the pore size distribution of the solid; because of the assumptions made, the accuracy of the method is severely limited, but it is useful for comparative purposes and is virtually the only method applicable in the pore size range of tens to hundreds of angstroms.

Chemisorption results from valency forces—from the sharing of electrons between the adsorbate molecule and the adsorbent—so that, in effect, a surface chemical compound is formed. Chemisorption is characterized by a high heat of adsorption (of the order of tens of kilocalories per mole, in contrast to the few kilocalories of physical adsorption) and by difficulty of reversal: to desorb a chemisorbed gas in a reasonable time requires a temperature much higher than that at which the chemisorption occurred. Even so the adsorbate may be released in a chemically changed form; thus carbon monoxide chemisorbed on zinc oxide at room temperature is desorbed as carbon dioxide at 300°C .

Chemisorption is an essential primary step in heterogeneous catalysis. At least one of the reactants must be chemisorbed on the surface of the catalyst, and each of its molecules then forms, on the surface, a "transition complex" with a chemisorbed molecule of the second reactant B, or with a molecule of B which hits it directly from the gas phase.

Physical adsorption is an extremely widespread phenomenon, frequently unwanted. The adsorption of water vapor by chemicals, by textiles, by building materials and by glass is frequently troublesome and can only be avoided by taking extreme precautions; sometimes, however, the adsorption of water may be beneficial, and it plays an important role, for instance, in the hygiene of clothing.

Adsorption, whether physical or chemical, also reduces the adhesion, and therefore the friction, between solids; gases can accordingly act as lubricants. In addition, adsorption diminishes the tensile strength of brittle solids; the breaking stress of glass when exposed to nearly saturated water vapor is four times less than when exposed to a vacuum. This is because of the part played by the adsorbed film at the tip of fine cracks in the brittle solid; the film reduces the free surface energy (in the thermodynamic sense) of the solid. Adsorption also causes a small (a fraction of 1 per cent) expansion of the solid, but the swelling pressure set up—i.e., the pressure which would have to be exerted on the solid to prevent expansion—is very high and may reach many atmospheres. Stresses set up in structures made up of

porous solids, such as cement and mortar, when they take up or lose vapors, particularly water, may be so great as to cause cracking.

Absorption is said to occur when the molecules of the gas or vapor actually penetrate into the solid phase itself, so that a solid solution is formed; hydrogen is absorbed by iron at elevated temperature in this way, and many synthetic polymers absorb water vapor; benzene vapor is extensively taken up by rubber and water vapor by gelatin. Extensive swelling occurs and if the solid is mechanically weak, the absorption may continue until the system becomes a liquid. An absorption isotherm (analogous to the *adsorption* isotherm discussed earlier) can be determined, but is generally complicated and is best handled theoretically as a branch of solution thermodynamics.

In *adsorption from solution*, when a solid having appreciable surface area (say ~ 1 square meter per gram) is shaken up with a solution of substance A in solvent B, both A and B are adsorbed, but to different relative extents. This manifests itself in a change in the composition, e.g., a change Δx_1 in the mole fraction of A in the solution. The problem is thus more complicated than in the adsorption of gases, and the measured isotherm—the curve of Δx_1 against x_1 —is not susceptible to any simple theoretical treatment. In a *dilute* solution of A, however, A is always relatively more adsorbed than the solvent B; and if A is colored, the resulting diminution in the concentration of A in the solution will be readily detected by eye or colorimetrically.

S. J. GREGG

References

- Gregg, S. J., "Surface Chemistry of Solids," London, Chapman & Hall, 1961.
 Adamson, A. W., "Physical Chemistry of Surfaces," New York, Interscience Publishers, 1960.
 Young, D. M., and Crowell, A. D., "Physical Adsorption of Gases," London, Butterworths, 1962.
 Hayward, D. O., and Trapnell, B. M. W., "Chemisorption," London, Butterworth's, 1964.

AERODYNAMICS

Aerodynamics is the science of the flow of air and/or of the motion of bodies through air. It is usually directed at achieving flow or flight with the maximum efficiency. Aerodynamics is a branch of Aeromechanics; the other main branch is *Aerostatics* (lift of balloons, etc.). In popular usage *Aerodynamics* differs from *Gasdynamics* in that the latter considers other gases and products of combustion (and combustion); from *Aerophysics* which implies a substantial meteorological contribution; and from *Hydrodynamics* which implies employing a medium of density approximating that of useful bodies in it, and not infrequently a sharp limit to its extent (i.e., a water surface).

Aerodynamics is conveniently divided into low and high speed regimes. (The latter in the

articles on COMPRESSIBILITY, FLUID DYNAMICS and SHOCK WAVES; here low speed aerodynamics is discussed.)

The many facets of aerodynamics include: (1) aerodynamic performance, (2) aerodynamic design, (3) aerodynamic loads, (4) aerodynamic structures, (5) aero-elasticity, (6) aerodynamic heating, (7) aerodynamic compressibility, and (8) aerodynamic research for all of the above.

The computation of aerodynamic effects are based on four laws (given as adapted for fluids):

(1) *Newton's Second Law*: "A force applied to a fluid results in an equal but opposite reaction which in turn causes a rate of change of momentum in the fluid."

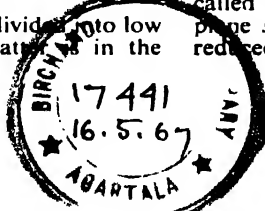
(2) *The Equation of State*: (also called the Gas Law): "The product of pressure and volume of a gas, divided by its absolute temperature, is a constant."

(3) *The Continuity Equation*: "The mass that passes a station in a duct, or in a natural tube bounded by streamlines in a given time, must equal that passing a second station in the same time."

(4) *The Energy Equation*: "The total energy in a mass of air remains constant unless heat or work is added or subtracted."

The above relations, written algebraically and limited or combined, yield a vast array of equations used to calculate the practical problems of aerodynamic. One of the *equations* thus derived is due to Bernoulli and is widely used in low speed aerodynamics. It states that in free flow the sum of the static and dynamic pressures is a constant. Static pressure is that pressure which is equal in all directions; dynamic pressure is the pressure rise realized by bringing the fluid to rest. Bernoulli's equation states that as air speeds up, its pressure falls, thus explaining the "lifting suction" as the airstream traverses the curved upper surface of a wing.

The overwhelmingly important *law* of low speed aerodynamics is that due to Newton (1, above). Thus a helicopter gets a lifting force by giving air a downward momentum. The wing of a flying airplane is always at an angle such that it deflects air downward. Birds fly by pushing air downward. Propellers and jet engines make a forward force ("thrust") by giving air a rearward momentum. In the above statements, it is much more accurate to use the expression "downward (or rearward) momentum," rather than "downward (or rearward) velocity," as less dense air at high altitudes produces smaller forces for comparable velocity changes. Forcing the air downward does not occur instantaneously; it takes place over the lifting surface. Indeed, the motion reaches ahead of the airplane so that some downward velocity occurs before the airplane arrives. Thus, airplanes (helicopters, birds, etc.) are always "flying uphill" which is another way of saying it takes a force to fly even when the air is considered frictionless. This force is used to push air *forward*, and is called "drag due to lift." It *increases* as an airplane *slows down* or *flies higher*, and it may be reduced (for low speed aircraft) by *increasing* the



wing span. Supersonic aircraft, operating under the same laws, but in a different manner, still have a drag due to lift which increases as above, but is reduced designwise by *reducing* the span. Hence the new airplanes whose wings crank back for high speed flight. (The "uphill" concept has an analogy in the rolling resistance of a wheel. The weight a wheel carries deflects the surface on which it rests so that it sits in a "gully." Either way it rolls, the path is uphill).

A second important phenomena (but not a "law" in the sense that it cannot be circumvented) is the manner in which air flows over say, an airplane surface. Away from the surface some freestream velocity exists. As the surface is approached the local velocity becomes less than freestream and finally becomes zero at the surface. The zone in which the air is appreciably slower is called the *boundary layer*. Slowing the air reduces its momentum, and by Newton's law (above), a force is produced which acts in a direction to slow the airplane. Like forward force needed to produce lift, this *frictional drag* force must have a forward force supplied from somewhere to balance it. For high speed aircraft a heating occurs in the boundary layer which requires additional force. Skin friction drag is *decreased* by reducing the amount of surface the air scrapes against (i.e., making the aircraft smaller) and by *flying slower or higher*—essentially the opposite of the actions which reduce the drag due to lift. Thus the science of aerodynamics seeks the most efficacious melding of the two types of losses. Factors which must be added include providing space for fuel and people, and enough wing to yield a reasonable landing speed. The drag due to lift and the drag due to friction are balanced by the forward thrust provided by the propeller or jet engine (which also operates by Newton's law). However, this is not enough. The distribution of lift on wing and tail must be so located that the aircraft is aerodynamically balanced. Like a child's swing, upon being disturbed it should tend to return to its original ("trimmed") condition.

In the above paragraphs the concern has been for aerodynamic efficiency through optimum design for optimum performance. After these have been achieved (by studies of previous designs and tests in *wind tunnels*) the aerodynamicist provides aerodynamic loads to which an aero-structural engineer must design. Aero-structural design is one of the most challenging of all design problems as the loads must be carried by minimum weight. Aeroplanes carry no "factor of safety" (sometimes called "factor of ignorance"). The aerodynamic-loads engineer furnishes the maximum air loads the aircraft is ever expected to see; the structure is designed to withstand these loads without being permanently bent. If it ever sees a greater load it will be bent or destroyed; there is about a 50 per cent difference between maximum no-permanent set load and catastrophic destruction, depending on the material of which the aircraft is constructed.

In an effort to keep aerodynamic structures light, they often become flexible, and, in turn,

become susceptible to flutter, a motion similar to that of a flag in a wind. The motion gets worse with speed. It is the job of the aero-elasticity engineer to assure that flutter will not occur, usually through a redistribution of internal weights (fuel, etc.) and, rarely, through strengthening the structure.

At the higher speeds, the performance, loads, structures and elasticity problems are greatly worsened by the aerodynamic heating which occurs. This is discussed in the article on COMPRESSIBILITY.

Aerodynamics is not only concerned with aircraft. The wind loads on signs, buildings, trees; the aerodynamics drag of autos, boats, trains; the air pollution from smoke stacks of factories and ships; the evaporation of open water; the blowing of sand and snow; and the internal losses of air-conditioning ducts—all deserve and are getting scrutiny by aerodynamicists. The forces on all are proportional to the rate of change of momentum they give, or are given by the air, and all have friction drag in their boundary layers. Aerodynamic research scientists seek to further understand and improve the air flow involved in each.

Aerostatics. Aerostatics is the science of making things (balloons, zeppelins, etc.) statically buoyant in the air. It is a branch of *Aeromechanics*; the other branch is *Aerodynamics*.

The basic principle of aerostatics is due to Archimedes "A body immersed in a fluid (or gas) is buoyed up by a force equal to the weight of the fluid (or gas) displaced." Thus for buoyancy the weight of structure plus the weight of the contained gas must equal the weight of the air displaced.

Wind Tunnels. Wind tunnels are devices which provide an airstream of known and steady conditions in which models requiring aerodynamic study are tested. The essential elements of a tunnel are

- (1) A drive system consisting of either a compressor for continuous operation or a tank of compressed air for intermittent operation;
- (2) A test section in which models are held and their orientation is changed;
- (3) Instrumentation capable of reading force, pressure, and optical effects produced by the model;
- (4) An air efflux system consisting of free exit to the atmosphere or to a vacuum tank, or a tunnel returning the air to the compressor.

There are approximately 500 wind tunnels in the country, ranging in test section size from 1 inch \times 1 inch to 30 feet \times 60 feet, with speeds from 50 to 7000 mph.

ALAN POPE

References

- Perkins, Courtland, D., and Hage, Robert E., "Airplane Performance, Stability, and Control," New York, John Wiley & Sons, Inc., 1949.
- Kuethe, A. M., and Shetzer, J. D., "Foundations of Aerodynamics," New York, John Wiley & Sons, Inc., 1959.

Pope, Alan, "Basic Wing and Airfoil Design," New York, McGraw-Hill Book Co., 1951.

Pope, Alan, "Wind Tunnel Testing," New York, John Wiley & Sons, Inc., 1954.

ALTERNATING CURRENTS

Definition of Alternating Current. An alternating current is a periodic function of the time, the function being such that the average value is zero. A special case of an alternating current is shown in Fig. 1. The square wave is clearly periodic, and in any one cycle, the area under the curve above the horizontal axis is equal to the area below the horizontal axis. If the two areas are not equal, the current may be described as an alternating current superposed on a direct current, provided the resultant current varies in a cyclic manner.

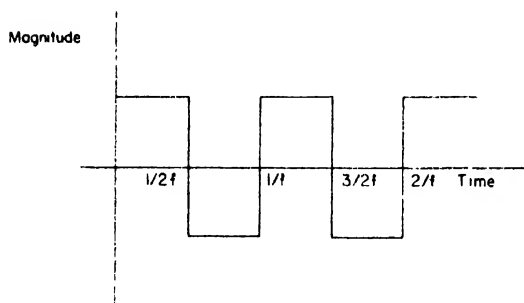


FIG. 1. A square wave alternating current.

In general, any alternating current may be considered to be the sum of a Fourier series of sinusoidal waves. For example, the square wave shown in Fig. 1 may be written as

$$\frac{4A}{\pi} (\sin 2\pi ft + \frac{1}{3} \sin 6\pi ft + \frac{1}{5} \sin 10\pi ft + \dots)$$

where A is the amplitude of the square wave, f is the frequency in cycles per second, and t is the time in seconds. Since any alternating current may be expressed as the sum of a series of sinusoidal terms, the remainder of this article will be devoted to a discussion of sinusoidal voltages and currents.

Root-mean-square Value. The equation for an alternating current i may be written as

$$i = I_m \sin(\omega t - \delta) \quad (1)$$

where I_m is the maximum or peak value of the current, ω is 2π times the frequency f in cycles per second, and δ is a phase angle. A graph of the current i is shown in Fig. 2. Since the positive and negative loops are mirror images, the average value of the current over a complete cycle is zero. The latter statement is valid for all alternating currents and, hence, gives no information about a particular alternating current.

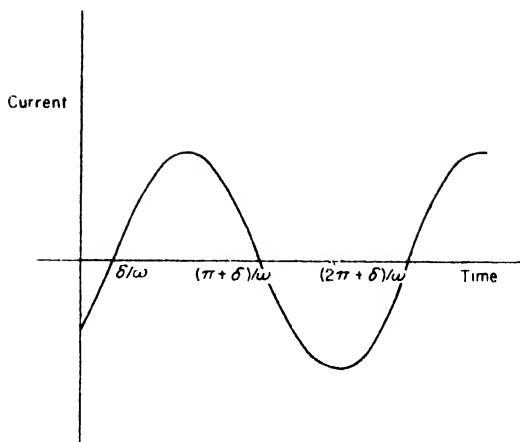


FIG. 2. A sinusoidal alternating current.

A useful way of stating the magnitude of an alternating current is to give its effective or root-mean-square value. The term root-mean-square is derived from the idea of taking the square root of an average square of the current. Thus, by definition, the effective value I_e of the current i given by Eq. (1) is

$$I_e = \sqrt{\frac{\omega}{2\pi} \int_0^{2\pi/\omega} I_m^2 \sin^2(\omega t - \delta) dt} \quad (2)$$

where $2\pi/\omega$ is the time for one cycle. In effect, the quantity under the square root sign is the sum of the squares of the currents during one cycle divided by the time for one cycle. The value of I_e may be found by performing the integration. The result is

$$I_e = \frac{I_m}{\sqrt{2}} = 0.707 I_m \quad (3)$$

Clearly, the effective value of a sinusoidal alternating current is 70.7 per cent of the maximum or peak value. Similarly the effective value of a sinusoidal voltage is 70.7 per cent of the maximum or peak value.

Alternating-current Series Circuit. A simple alternating-current series circuit is shown in Fig. 3. At the instant considered in the diagram, the current is in the direction shown. The circuit consists of a generator connected in series to a pure resistance R , a pure inductance L , and a capacitance C . It is important to understand the relationship of the current to the potential difference across each element. In each case, the best starting point is a basic definition. According to Ohm's law for a pure resistance,

$$R = \frac{V_R}{i} \quad (4)$$

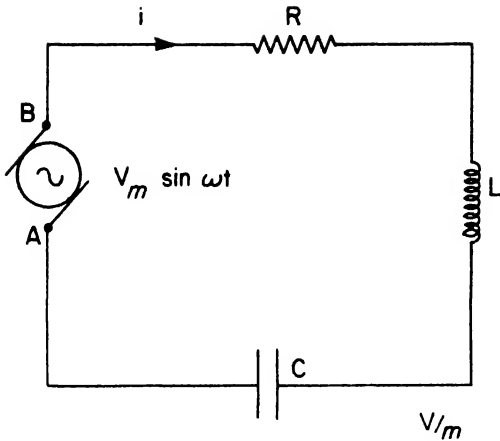


FIG. 3. An ac series circuit.

where V_R is the voltage drop across the resistance. The inductance L of a coil is given by

$$L = \frac{N\phi}{i} \quad (5)$$

where N is the number of turns and ϕ is the magnetic flux passing through one of the turns of the coil as a result of the current i . Writing Eq. (5) as

$$Li = N\phi$$

and then differentiating both sides, we obtain

$$L \frac{di}{dt} = N \frac{d\phi}{dt} \quad (6)$$

According to Faraday's law, the right-hand side of Eq. (6) is the magnitude of the induced emf. The left-hand side, $L di/dt$, therefore, is the voltage drop V_L across the inductance L . Finally, the capacitance C is by definition

$$C = \frac{q}{V_C} \quad (7)$$

where q is the instantaneous charge on the positive plate and V_C is the drop in potential in going from the positive plate to the negative plate.

The relation between the impressed voltage $V_m \sin \omega t$ and the instantaneous current i follows from Kirchhoff's law that the sum of the differences in potential in going around a complete circuit must be zero. At the instant shown in Fig. 3, there is a potential rise $V_m \sin \omega t$ in going from A to B and there are potential drops, V_R , V_L , and V_C in traversing the rest of the circuit. According to Kirchhoff's law

$$V_m \sin \omega t - V_R - V_L - V_C = 0$$

$$V_m \sin \omega t - Ri - L \frac{di}{dt} - \frac{q}{C} = 0$$

Since the current is the rate of flow of charge

$$i = \frac{dq}{dt} \quad (8)$$

It is now possible to express the current i as a function of t . The result neglecting initial transient effects is

$$i = \frac{V_m \sin(\omega t - \delta)}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}} \quad (9)$$

where $\tan \delta = (\omega L - 1/\omega C)/R$. The phase angle δ is the angle by which the current i lags behind the impressed voltage.

The maximum or peak value of i is

$$I_m = \frac{V_m}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}}$$

If both sides of this equation are divided by $\sqrt{2}$, we obtain

$$\begin{aligned} \frac{I_m}{\sqrt{2}} &= \frac{V_m/\sqrt{2}}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}} \\ I_e &= \frac{V_e}{\sqrt{R^2 + (\omega L - 1/\omega C)^2}} \end{aligned} \quad (10)$$

Equation (10) states the relation between the effective value or the current and the effective value of the impressed voltage.

Impedance and Reactance. The denominator of Eq. (10) may be defined as the impedance Z of the circuit. We may therefore write

$$I_e = \frac{V_e}{Z} \quad (11)$$

Equation (11) is similar in form to Ohm's law, Eq. (4). The impedance Z is the square root of the sum of the squares of two terms. The first is the resistance R , and the second is $\omega L - 1/\omega C$. The latter is called the reactance. The quantity ωL is the inductive reactance whereas $1/\omega C$ is the capacitive reactance. If the inductive reactance is greater than the capacitive reactance, the phase angle δ is positive and the current lags behind the voltage. If the inductive reactance is less than the capacitive reactance, the current leads the voltage.

Vector Diagram. The current and the various voltages may be related in a meaningful way by means of a vector diagram. Equation (10) may be rewritten as follows:

$$V_e = \sqrt{R^2 I_e^2 + (\omega L I_e - I_e/\omega C)^2} \quad (12)$$

Equation (12) implies that V_e is the resultant of a vector RI_e at right angles to a vector $\omega L I_e - I_e/\omega C$. This is shown in Fig. 4. The current I_e is drawn along the horizontal axis, and the effective voltage drop across the resistance, RI_e , is also drawn along this axis. The effective voltage drop across the coil is $\omega L I_e$, and this potential difference is drawn along the positive vertical

axis. Finally, the effective potential drop across the capacitor is $I_e/\omega C$, and this vector is drawn along the negative vertical axis. The resultant of the three vectors is V_e , in agreement with Eq. (12).

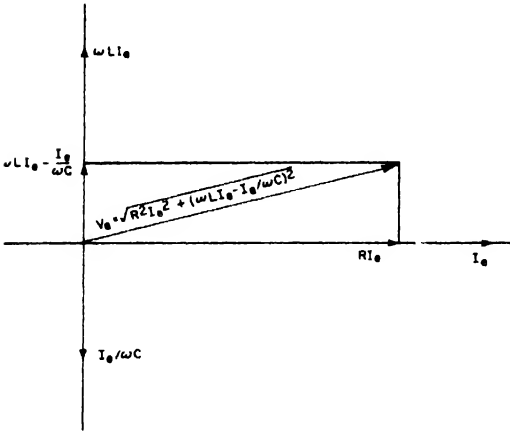


FIG. 4. A vector diagram for an ac series circuit

Resonance. If the capacitance C can be varied, the current I_e will be a function of C in accordance with Eq. (10). When

$$\omega L = \frac{1}{\omega C} \quad (13)$$

the effective current will be a maximum. The circuit is then said to be in resonance. Actually, the inductance L or the angular frequency ω may be varied instead of the capacitance C . The circuit will be in resonance whenever Eq. (13) holds. At resonance, the impedance Z is equal to R , and the circuit, under such circumstances, acts as though it contains resistance only. The process of obtaining resonance is called tuning the circuit.

Average Power. The potential difference V across an ac generator at any instant is the work required to transfer a unit charge from the negative to the positive terminal. The work done in transferring a charge dq is consequently $V dq$, and the work done per unit time is

$$P = \frac{V dq}{dt} \quad (14)$$

where P is, by definition, the instantaneous power and dt is the time interval to transfer the charge dq . Since

$$i = \frac{dq}{dt}$$

Eq. (14) may be written

$$P = Vi$$

The instantaneous power is the product of the instantaneous voltage and current.

When alternating current circuits are con-

sidered, the average power \bar{P} rather than the instantaneous power is of interest. By definition

$$\bar{P} = \frac{\omega}{2\pi} \int_0^{2\pi/\omega} Vi dt \quad (15)$$

where, as before, $2\pi/\omega$ is the time for a complete cycle. The average power may be evaluated by making the following substitutions in Eq. (15):

$$V = V_m \sin \omega t$$

$$i = I_m \sin (\omega t - \delta)$$

The result is

$$\bar{P} = \frac{1}{2} V_m I_m \cos \delta \quad (16)$$

Equation (16) may be rewritten

$$\begin{aligned} \bar{P} &= \frac{V_m}{\sqrt{2}} \frac{I}{\sqrt{2}} \cos \delta \\ \bar{P} &= V_e I_e \cos \delta \end{aligned} \quad (17)$$

Evidently, the average power is the effective voltage times the effective current multiplied by the cosine of the phase angle. In this connection, $\cos \delta$ is called the power factor. Equation (17) may be interpreted to mean that only the component of V_e in phase with I_e contributes to the average power. The other component may be said to be wattless. Since

$$V_e = I_e Z$$

and

$$\cos \delta = \frac{R}{Z}$$

Eq. (17) may be written as follows:

$$\bar{P} = I_e^2 R \quad (18)$$

From the latter form, it may be concluded that the average power is the average rate at which heat is developed in the circuit. Equation (18) also shows that a direct current having a value I_e would produce the same heating effect as an alternating current having an effective value I_e .

The Complex-number Method. In the foregoing, an alternating-current series circuit was discussed by representing voltages as vectors in the real plane. For more complicated circuits, this method is too clumsy. It is much more convenient to deal with vectors analytically by utilizing the i -operator. By definition,

$$j = \sqrt{-1}$$

When a real number is multiplied by j , it becomes an imaginary number. In other words, a point on the real axis is rotated through 90 degrees so that it becomes a point on the imaginary axis. The "complex" impedance of a series circuit may thus be written

$$Z = R + j\left(\omega L - \frac{1}{\omega C}\right) \quad (19)$$

since the reactance may be considered to be at right angles to the resistance. When several impedances are connected in series, the total complex impedance is

$$Z = Z_1 + Z_2 + Z_3 + \cdots \quad (20)$$

and, when several impedances are connected in parallel, the total impedance is given by

$$\frac{1}{Z} = \frac{1}{Z_1} + \frac{1}{Z_2} + \frac{1}{Z_3} + \cdots \quad (21)$$

The effective voltage V across the generator may be considered to be a vector along the real axis. The effective current I furnished by the generator is therefore

$$I = \frac{V}{Z} \quad (22)$$

By solving Eq. (22), the magnitude of I and the phase relation between I and V may be found. Although new mathematical techniques are needed, the saving of time usually justifies the use of the complex number method of handling complicated ac circuits.

REUBEN BENUMOF

References

- Benumof, Reuben, "Concepts in Electricity and Magnetism," Ch. 14, New York, Holt, Rinehart, and Winston, 1961.
- Benumof, Reuben, "Concepts in Physics," Ch. 14, Englewood Cliffs, N.J., Prentice-Hall, Inc., 1965.
- Frank, N. H., "Introduction to Electricity and Optics," Ch. 9, New York, McGraw-Hill Book Co., 1950.
- Peck, E. R., "Electricity and Magnetism," Ch. 11, New York, McGraw-Hill Book Co., 1953.
- Duckworth, H. F., "Electricity and Magnetism," Ch. 13, New York, Holt, Rinehart, and Winston, 1960.
- Scott, W. T., "The Physics of Electricity and Magnetism," Ch. 9, New York, John Wiley & Sons, 1959.

Cross-references: CIRCUITRY, ELECTRICITY, POTENTIAL, RESONANCE.

ANGULAR MOMENTUM. *See* ROTATION—CIRCULAR MOTION.

ANTENNAS

Communication systems, characteristically, consist of cascaded networks, each network designed to carry out some operation on the energy conveying the information. In radio communication systems, antennas are the networks serving to transfer the signal energy from circuits to space and, conversely, from space to circuits. In circuits, the flow of energy is restricted to one or the other of two directions. The effectiveness of transfer of energy between the antenna

and the adjacent circuit element is, therefore, determined solely by the terminal impedance of the antenna and that of the adjacent circuit. The knowledge of the antenna terminal impedance over the desired frequency range, therefore, fully describes the joint performance of the antenna and the circuit element.

The relationship between the antenna and space, however, is much more complex. The distribution of the radiated energy varies with the direction in space and with the distance from the antenna. This gives rise to the directive properties of the antenna. Further, the energy is radiated in the form of an electric and a magnetic field. These are vector quantities which, at a distance from the source, are at right angles to each other and to the direction of propagation. The planes in which these vectors are located, and whether they are stationary or rotate with time, determine the polarization of the radiated field. The performance of an antenna can, therefore, be fully described only by specifying several parameters, such as radiation pattern, gain, and polarization. It is convenient, in discussing antenna properties, to consider the antenna as a radiating rather than a receiving network. The antennas are, however, linear networks and are subject to the law of reciprocity.¹ The performance of an antenna, therefore, in terms of radiation pattern, gain, or polarization is the same, irrespective of whether the antenna radiates or absorbs radiation.

Except for the immediate neighborhood of the antenna, referred to as the "near-field region" of the antenna, radiated energy propagates radially from the antenna, and the radiation intensity* varies inversely as the square of the distance from the antenna. This is a propagation effect. In discussing the antenna performance, it is customary to disregard this and to represent the distribution of the radiated power as a function of the two direction angles only. Such a distribution is commonly represented graphically and is then known as the radiation pattern of the antenna. The radiation patterns can take a variety of forms. Sometimes they are in the form of a polar diagram, with the radial distance proportional to either field strength* or intensity. The intensity may be represented linearly, as power, or logarithmically, in decibels.² For representing the directive properties of an antenna in all directions, contours of equal radiation intensity may be plotted, with the two direction angles as abscissas and ordinates, respectively.

The directive properties of an antenna also lead to the concept of antenna "gain." The directive gain of an antenna in a specified direction is the radiation intensity in that direction compared to what it would be if the total radiated power were distributed equally in all directions. For some applications, such as point-to-point communication, high values of antenna gain are

* In this discussion, radiation intensity has the dimensions of power flow per unit area, normally, watts per square meter. Electric field strength, on the other hand, is in volts per meter.

desired because such antennas concentrate the available power, thus effectively increasing it. Conversely, in receiving applications, such antennas are more responsive to radiation arriving from one direction. For other applications such as broadcasting, antennas with low directivity may be desired.

The gain of an antenna is dependent principally upon the size of the antenna, expressed in wavelengths. The larger the antenna, the greater is likely to be its gain. The values of gain for different antennas range from 1.5 for an electrically small dipole to hundreds and even thousands times that. In practice, antenna gains are usually expressed logarithmically, in decibels. For the low-frequency end of the radio spectrum (15 kc/sec to 3 Mc/sec), antennas, although large physically, are relatively small in terms of wavelengths. Therefore, the directive gains of these antennas seldom exceed 3 (4.8 db). In the high-frequency band (3 to 30 Mc/sec), which is used principally for long-distance communication, antenna gains of 10 to 100 (10 to 20 db) are frequently encountered. At microwave frequencies, where the wavelengths are a fraction of a meter, gains of several hundred, and even thousand times (20 to over 30 db), are common.

When an antenna has one or more of its dimensions significantly larger than a wavelength, its radiation pattern is likely to have more than one maximum. The radiation pattern, in such cases, is said to have a lobe structure. That part of the radiation pattern which encompasses the direction of the largest maximum and the radiation immediately to each side of it, is referred to the main lobe. The radiation about the minor maxima is referred to as the secondary or side lobes. One of the frequent goals in antenna design is the reduction in the levels of secondary lobes. These may, at times, be a source of interference to other transmissions.

In common with light, radio waves consist of electric and magnetic fields at right angles to each other and to the direction of propagation. The orientation of these fields relative to the observer determines the polarization of the wave. In radio terminology, the orientation of the electric vector of a radio wave is taken as the direction of polarization. Thus, if the electric field vector is parallel to the ground, the radio wave is termed "horizontally polarized." Although the polarization of the energy radiated by an antenna, in general, varies with the direction, an antenna is usually designated as being horizontally (or vertically, or circularly, etc.) polarized, depending on the polarization of its radiation in the direction of the main lobe maximum.

The importance of polarization in radio engineering lies principally in the different reflective properties of the ground for waves with electric field parallel to the ground and those normal to the ground. Different radio services are served best by different polarizations. Antennas for use in the low-frequency end of the radio spectrum, 15 kc/sec through about 3 Mc/sec, are almost invariably vertically polarized. This includes the

AM broadcast band. In the high-frequency band, 3 to 30 Mc/sec, both horizontal and vertical polarizations are used. For FM and television broadcast service in the United States and many other countries (but not in the United Kingdom), horizontal polarization is employed.

The types and variations of antennas encountered in practice are extremely numerous. Each type has some advantage over the others for some specific requirement. Among some of the more important and frequently encountered requirements are those for operating bandwidth, high radiation efficiency, specified degree of directivity, whether high or low, and polarization. A few of the representative types of antennas frequently encountered in practice are illustrated in Fig. 1, 2, and 3. Figure 1 shows two of the

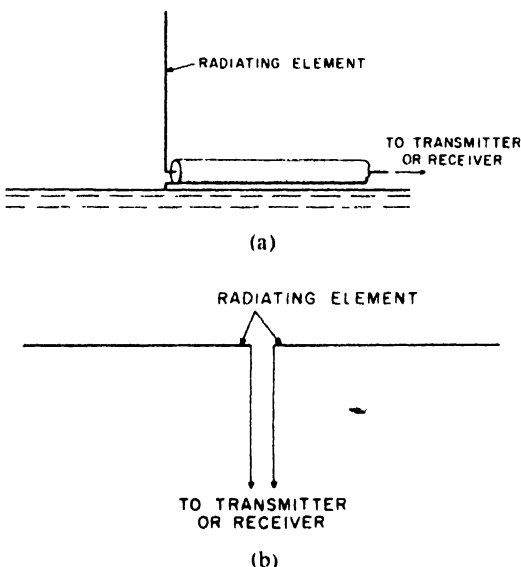


Fig. 1. Two types of elementary radiators: (a) Monopole over ground. (b) A dipole.

elementary types of radiators, a monopole and a dipole. A monopole, shown in Fig. 1(a), in one form or another, is employed almost exclusively throughout the low-frequency end of the radio spectrum. The dipole, shown in Fig. 1(b), is somewhat more versatile, as it can be oriented to give either horizontal or vertical polarization. It is frequently used as an elementary radiator in large array-type antennas.

Figure 2 presents two, highly directive, but otherwise radically different, types of antenna. The rhombic antenna shown in (a) has broadband properties and is used widely in point-to-point communication service. The Yagi antenna displayed in (b) illustrates a relatively compact antenna with high gain for its size. Its operating frequency band is quite narrow.

The last figure, Fig. 3, shows two antenna types frequently used at microwave frequencies. The horn antenna is used generally where moderate directivity suffices. The parabolic antenna, on

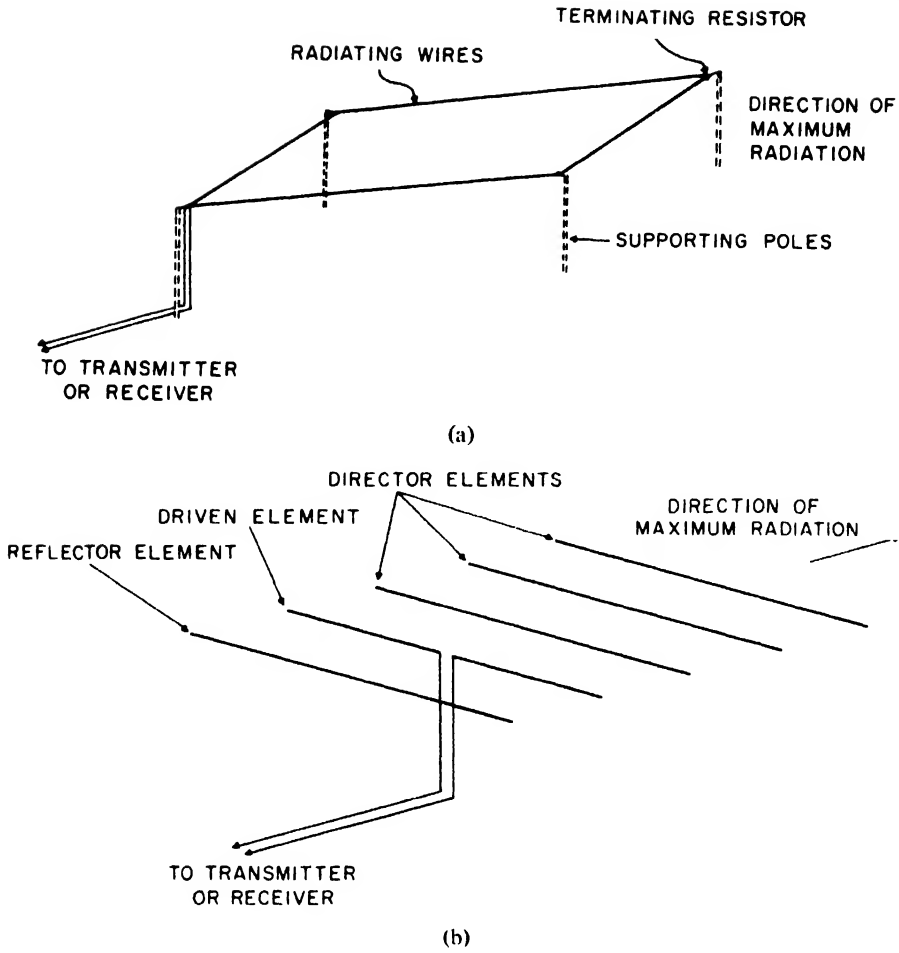


FIG. 2. Examples of directive antennas: (a) Rhombic antenna. (b) Yagi antenna.

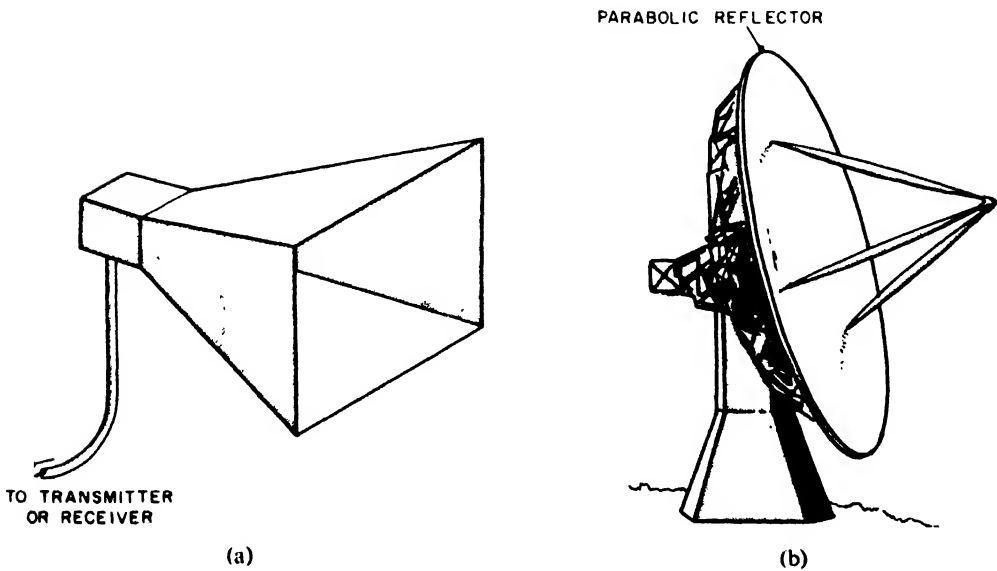


FIG. 3. Microwave-type antennas: (a) Horn antenna. (b) Parabolic-reflector antenna.

the other hand, is used for high-gain applications and is a quasi-optical device.

For a more complete listing of antenna types and their discussion, the reader is referred to texts on antennas such as the "Antenna Engineering Handbook."³ For additional discussion of the principles underlying antennas, texts by Kraus⁴ and Schelkunoff and Friis⁵ are suggested.

H. V. COTTONY

References

1. Jordan, E. C., "Electromagnetic Waves and Radiating Systems," pp. 327-328, Englewood Cliffs, N.J., Prentice Hall, Inc., 1950.
2. Schelkunoff, S. A., "Electromagnetic Waves," pp. 25-26, New York, D. Van Nostrand Co., 1943.
3. Jasik, H., Ed., "Antenna Engineering Handbook," New York, McGraw-Hill Book Co., 1961.
4. Kraus, J. D., "Antennas," New York, McGraw-Hill Book Co., 1950.
5. Schelkunoff, S. A., and Friis, H. T., "Antennas, Theory and Practice," New York, John Wiley & Sons, 1952.

Cross-references: CIRCUITRY, PROPAGATION OF ELECTROMAGNETIC WAVES, WAVEGUIDES.

ANTIFERROMAGNETISM

Antiferromagnetism is the most common form of magnetic order. It is found in almost all inorganic compounds of the transition metals, rare earths and actinide elements; it is also found in Cr, Mn, Pt, Pd and rare earth metals and alloys, although the situation is rather more complicated in the case of metals, and the discussion here will be most appropriate to insulators. The principal feature of antiferromagnetism is the spontaneous antiparallel alignment of neighboring electron spins which takes over from the paramagnetic state. The critical temperature of this second-order phase transition is called the Néel temperature (T_N). The strength of the ordering interaction is characterized by the magnitude of T_N which ranges from below 1°K to above room temperature. A few antiferromagnets whose properties have been studied in detail are listed along with their Néel temperatures: NiO, 520°K; Cr₂O₃, 310°K; MnF₂, 67.4°K; CuCl₂·2H₂O, 4.3°K.

The ordering interaction between neighboring metal spins in an insulator is called superexchange since it takes place via an intervening anion, O, F, S, etc. (cf. direct exchange between adjacent ions as encountered in FERROMAGNETISM). Superexchange results from charge transfer (see BOND, CHEMICAL), and it is best illustrated by a typical example, say MnO.

Ground and Excited States of (MnOMn)⁺⁺



In the ground state, the purely ionic configuration, there is no interaction between metal ions. If,

however, one of the two bonding electrons of O⁻ is transferred to the Mn⁺⁺ at left, there will be strong Hund's rule coupling within that ion, and also, the unpaired electron on O⁻ can couple with the Mn⁺⁺ at right. Since the two bonding electrons on O⁻ have opposite spins, the overall interaction will appear as antiparallel exchange coupling between the two Mn ions.

As with ferromagnetism, the magnetic properties of the exchange interaction can be described by assuming an energy in the form

$$\mathcal{H} = \sum J_{ij} S_i \cdot S_j$$

where S_i and S_j are the spin angular momentum vectors of a pair of ions. The dominant exchange constant is between near-neighbor pairs and is negative in sign, so that antiparallel arrangement results. The magnetic structure of an antiferromagnet can be decomposed into two interpenetrating sublattices having oppositely directed magnetizations, M^+ and M^- . The exchange interaction can then be simply represented in terms of the sublattice magnetizations by defining an exchange field proportional to the sublattice magnetization and acting on one sublattice in a direction opposite that of the other:

$$H_E^+ = -\lambda M^-$$

The exchange parameter λ is nearly temperature independent, and the sublattice magnetization grows from zero at T_N to 100 per cent aligned at low temperatures in exactly the same way that the saturation magnetization of a ferromagnet behaves below T_c . Also like ferromagnetism, spin wave theory describes low-lying fluctuations from alignment.

While the exchange interaction produces antiparallel alignment of the spins, their direction with respect to the crystalline axes is a consequence of the magnetic anisotropy, the energetic inequivalence of direction. There are three origins of anisotropy: (1) dipole-dipole interaction among the array of atomic moments which gives anisotropy in all but cubic symmetry; (2) Stark-effect interaction of each single ion with the local crystalline electric fields; and (3) anisotropic exchange, a result of spin-orbit coupling and isotropic superexchange between excited orbital states. The latter two mechanisms are important for non-S-state ions, especially in crystals of low symmetry. The anisotropy energy, $K(\alpha, \beta, \gamma)$, is expressed as a function of the direction cosines of the sublattice magnetizations. When the sublattice is close to the equilibrium direction, called the easy axis, an effective field along the easy axis called the anisotropy field, H_A^+ , is defined in such a way as to produce the appropriate restoring torque on M^+ for small deviations from the easy axis.

The vector model, which uses exchange and anisotropy fields, is very useful in explaining the major experimental features of antiferromagnetism. It is used in Fig. 1(a) to calculate the perpendicular static susceptibility of a typical antiferromagnet, which is shown in Fig. 1(b). At

absolute zero, the susceptibility parallel to the easy axis is zero, since no spin is able to turn over against the exchange field; thermal fluctuations permit an increasing susceptibility with increasing temperature. The dynamic susceptibility may also be derived from the vector model by writing

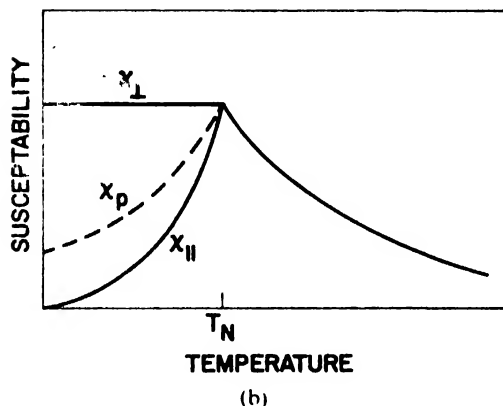
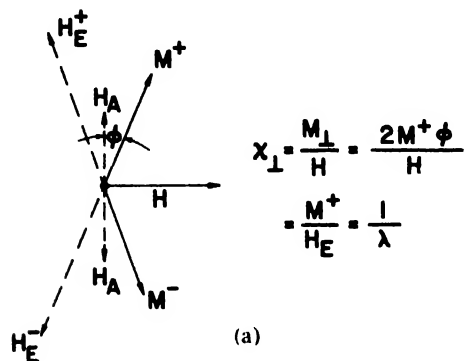


FIG. 1. (a) The vector model of a simple antiferromagnet is used to calculate the susceptibility to an applied field perpendicular to the easy axis. (b) The susceptibility of a typical antiferromagnet. X_D is the spherical average of the parallel and perpendicular susceptibilities that is observed in a powder specimen.

the Bloch equations for both sublattices, including the approximate effective fields. Antiferromagnetic resonance occurs at a frequency $\omega \approx \gamma \sqrt{2H_E H_A}$ where γ is the electron gyro-magnetic ratio, 2.8 Mc/oersted. Antiferromagnetic resonance has been observed in several materials, often being found at submillimeter wavelengths. In addition to static susceptibility and antiferromagnetic resonance, the principal experiments in antiferromagnetism have been neutron diffraction, which yields the spin structure and sublattice magnetization and sometimes even more details, and nuclear magnetic resonance, which yields high-precision measurements of the sublattice magnetization and insight into the details of charge distributions around ions.

DALE T. TEANEY

References

- Nagamiya, T., Yosida, K., and Kubo, R., "Antiferromagnetism," in *Advan. Phys.*, 4, 1 (1955).
Rado, G. T., and Suhl, H., Eds., "Magnetism," New York, Academic Press, 1963.

Cross-references: BOND, CHEMICAL; FERRIMAGNETISM; FERROMAGNETISM; MAGNETISM; MAGNETOMETRY; PARAMAGNETISM.

ANTIPARTICLES

One of the great discoveries of modern physics is that for every one of the elementary entities of matter and radiation—"particle"—there exists a corresponding entity—"antiparticle"—with certain of the particle-defining properties identical and others reversed in sign. The properties identical for particle and antiparticle are those determined by space-time symmetry: spin s , mass m , and lifetime τ . Those which are reversed are the internal-symmetry properties: electric charge Q , baryon charge B , light or heavy lepton charge l or L . In contrast with mass and spin, the internal symmetry properties have values which are simply additive in a composite system, and which are conserved in any process; their sign reversal for antiparticles is therefore essential to maintaining the conservation law in the dramatic processes of *pair creation* and *pair annihilation* in which antiparticles are observed appearing and disappearing with their conjugate particles. Also opposite for particle and antiparticle is the approximately conserved internal symmetry property hypercharge Y (the sum of the algebraically largest and algebraically smallest of the electric charges in the mass multiplet to which the particle belongs). Only those particles for which the values of all internal properties are zero, e.g., the photon and the neutral pion, are identical with their antiparticles (see CONSERVATION LAWS AND SYMMETRY).

The progressive recognition of the existence of antiparticles was initiated by Dirac's relativistic anti-electron theory in 1931 and by Anderson's independent experimental discovery of the anti-electron (positron) in 1932. An enumeration of known particles and antiparticles up to 1964 will be found under ELEMENTARY PARTICLES. Here we limit ourselves to a brief account of the theory, particularly of the particle-antiparticle conjugation operator C . This conjugation operator is known to be a symmetry (invariance) operator not only for the energy-momentum (and therefore mass) of free stable particles, but also for the total energy-momentum of a system of particles interacting according to strong or electromagnetic interactions (C invariance). In weak interactions (e.g., non-electromagnetic radioactive decay), it has been found that C invariance by itself does not hold, nor does space-inversion symmetry; only the product of the two (CP invariance) holds approximately. Recent experiments show a small nonvanishing probability for decay of the second type neutral kaon into two pions; this points to a

slight violation also of CP invariance in hypercharge-changing weak interactions.

The best (though not completely adequate) theory that we have of elementary particles and antiparticles is *relativistic quantum field theory*. "Relativistic" means that we require the space-time symmetry of *special relativity*. The representation theory of the relativity group prescribes the characterizing properties of spin and mass. However, while the specific range of observed spin values (all half-integral multiples of a fundamental unit) is explained, the specific range of observed masses is as yet not fundamentally interpreted.

Besides the requirements of relativity, the theory is a QUANTUM THEORY in that it recognizes that the particle-characterizing properties listed above—and related *state-characterizing properties* such as energy-momentum P and angular momentum J —are not all mutually "sharp" or "compatible" observables; compatible observables are such as exhibit in a given system in appropriate states, a characteristic definite value ("quantum number") upon consecutive repeated measurements in any order. A complete set of compatible observables (C.S.C.O.) is any set needed to define the states uniquely, and the individual state is labeled by the set of values which the C.S.C.O. takes on for it. The states are represented as vectors (more exactly "rays") in a Hilbert space and the observables as linear operators. Compatibility (noncompatibility) of two observables is represented by commutativity (noncommutativity) of the corresponding two operators for which we use the same symbols as for the observables. For instance we have:

$$Cm = mC = 0 \quad (CQ = QC \neq 0) \quad (1)$$

In many instances two such operators anticommute, and this is how quantum theory describes the reversals of sign of certain properties under C as indicated above. Thus we have

$$\begin{aligned} CQ + QC = 0 \quad \text{or} \quad CQC^{-1} = -Q \\ CB + BC = 0 \quad \text{or} \quad CBC^{-1} = -B, \text{ etc.} \end{aligned} \quad (2)$$

where the similarity transformation on the left of the second equation on each line represents the effect of particle conjugation on the corresponding charge operator.

To derive these commutation and anticommutation relations for composite systems when they are given for individual particles (by postulate—in accord with the indications of experiment!), and also to describe the relations between reactions in which particles and antiparticles are respectively involved, we adopt the postulates of Quantum Field Theory. This theory, necessarily a many-particle theory, is best described at first for the non-interacting case when we can take the basic states to be n -particle states, n having all possible values. A specific n -particle state $|n_1, n_2, n_3, \dots\rangle$ is characterized by the set of *occupation numbers* n_1, n_2, \dots ($n_1 + n_2 + \dots = n$) specifying the numbers of particles having quantum numbers q_1, q_2, q_3, \dots corresponding to a

(C.S.C.O.): Q_1, Q_2, Q_3, \dots . The state $|0, \dots, 0, \dots\rangle \equiv |\Omega\rangle$ with all $n_i = 0$ is called the *vacuum state*. As in any vector space, there is also the *zero vector* 0. The single-step destruction operator a_λ for particles, and creation operator b_λ^+ for antiparticles associated with the λ th quantum number are introduced in the usual way (see FIELD THEORY) such that they satisfy the standard commutation (anticommutation) rules if the particles satisfy Einstein-Bose (Fermi-Dirac) statistics. One then introduces the *field operator*, which is a linear expansion in the a_λ and b_λ^+ (all λ), and which formally satisfies certain field equations, the latter usually being chosen on the basis of a classical analogy or on the basis of relativistic covariance and general agreement with experiment. All of the consequences of interest then follow from the definition of the field operator and from that of the particle conjugation operator:

$$\begin{aligned} Ca_\lambda C^{-1} &= \eta_c a_\lambda & Ca_\lambda^+ C^{-1} &= \eta_c^* b_\lambda^+ \\ Cb_\lambda C^{-1} &= \eta_c b_\lambda^+ & Cb_\lambda^+ C^{-1} &= \eta_c^* a_\lambda \end{aligned} \quad (3)$$

In Eq. (3) η_c is a phase constant ($|\eta_c| = 1$) which is nonmeasurable and can be chosen ± 1 by convention. If particle and antiparticle are identical—as is true for the photon and neutral pion alone among the "elementary particles" (but also true for certain composites like the singlet positronium when considered as a "particle")—then $b_\lambda = a_\lambda$, and η_c becomes measurable and equal to ± 1 , two physically distinct cases.

For a specific theory, particularly if it is given in Lagrangian form (in which case each of the observables listed above is defined by a symmetry operation on the Lagrangian), these observables can be written as bilinear functionals in the field operators.

As an example, we consider the case of a scalar (spin zero) charged field $\Phi(x)$ associated with a mass m , and satisfying the Klein-Gordon equation

$$(\square + m^2)\Phi(x) = 0 \quad (4)$$

For the C.S.C.O. we choose the linear momentum and the (not-independent) energy. Denoting the destruction operator for the particle of momentum k and energy $k_0 = \sqrt{k^2 + m^2}$ by a_k , and the creation operator for the corresponding antiparticle by b_k^+ , the field operator is given by

$$\Phi(x) = \frac{1}{\sqrt{2(2\pi)^3}} \int \frac{d^3k}{k_0} (a_k e^{-ikx} + b_k^+ e^{ikx}) \quad (5)$$

An explicit representation for C is

$$C = \exp \left[\frac{\pi i}{2} \sum_k (a_k^+ \eta_c^* b_k^+)(a_k - \eta_c b_k) \right] \quad (6)$$

and for the total energy-momentum operator, obtained for instance from the Lagrangian for Eq. (4), we have

$$P_\mu = \int \frac{d^3k}{k_0} k_\mu (a_k^+ a_k + b_k^+ b_k) \quad (7)$$

It is now possible by direct calculation to establish whether P_μ (or any other observable) commutes or anticommutes with C . In this way we can derive for all the free physical fields which have been considered applicable to nature, the results stated earlier in this article concerning preservation and reversal of signs of the fundamental quantum numbers.

It is possible in a similar way to examine the validity of invariance under particle conjugation of the interactions between fields. For instance, the electromagnetic interaction between the electron-field current j_μ and the photon-field potential A_μ is $j_\mu A_\mu$. The two results found for the separate fields by the methods described in the foregoing,

$$Cj_\mu C^{-1} = -j_\mu \quad CA_\mu C^{-1} = A_\mu$$

then guarantee that

$$Cj_\mu A_\mu C^{-1} = j_\mu A_\mu$$

Similarly, C -invariance holds for the accepted forms of strong interactions.

C -invariance does not hold for the V - A (polar vector minus axial vector) four-fermion type of weak interaction occurring, for example, in β -radioactivity. It has been shown however that for quite general interactions with proper orthochronous space-time symmetry, provided they are local, there is automatically invariance under the combined operation CPT where T represents time inversion (CPT Theorem). "Proper orthochronous" space-time symmetry of the interactions means that they are invariant under rotations, translations in space and time, and shifts to uniformly moving frames: "locality" means that the interactions consist of a linear combination of products of the fields and finite-order derivatives of the fields. Equality of mass and lifetime, anti-equality of charge and magnetic moment, and (with certain restrictions) conjugacy of decay schemes for a particle and its antiparticle all follow from CPT invariance alone. And this rests only on proper space-time symmetry, general principles of quantum theory, and the locality requirement.

M. A. MELVIN

References

- Dirac, P. A. M., "Quantized Singularities in the Electromagnetic Field," *Proc. Roy. Soc. London, Ser. A*, **133**, 60 (1931).
 Wolfenstein, L., and Ravenhall, D.G., "Some Consequences of Invariance under Charge Conjugation," *Phys. Rev.*, **88**, 279 (1952).
 Lee, T. D., and Yang, C. N., "Elementary Particles and Weak Interactions," Office of Technical Services, Department of Commerce, Washington, D.C. 1957.
 Lüders, G., and Zumino, B., "Some Consequences of TCP-invariance," *Phys. Rev.*, **106**, 385 (1957).
 Wick, G. C., "Invariance Principles of Nuclear Physics," *Ann. Rev. Nucl. Sci.*, **9** (1959).

Segré, E., "Antinucleons," *Ann. Rev. Nucl. Sci.*, **9** (1959).

Feinberg, G., and Weinberg, S., "On the Phase Factors in Inversions," *Nuovo Cimento*, **14**, 571 (1959).

Melvin, M. A., "Elementary Particles and Symmetry Principles," *Rev. Mod. Phys.*, **32**, 477 (1960).

McConnell, J., "Theory of Antinucleons," *Progr. Elem. Particle Cosmic Ray Phys.*, **5** (1960).

Sachs, R. G., "CP Violation in K^0 Decays," *Phys. Rev. Letters*, **13**, 286 (1964).

Cross-references: BOSE-EINSTEIN STATISTICS AND BOSONS, CONSERVATION LAWS AND SYMMETRY, ELEMENTARY PARTICLES, FIELD THEORY, QUANTUM THEORY, RELATIVITY.

ARCHITECTURAL ACOUSTICS

Although the practice of architectural acoustics involves a wide variety of special problems and techniques, the basic reasons for acoustical design are simply,

- (a) to provide a satisfactory acoustical environment, not too noisy and often not too quiet, for people at work and relaxation;
- (b) to provide good hearing conditions for speech; and
- (c) to provide a pleasant acoustical environment for listening to music.

Designing for Satisfactory Acoustical Environment. Each acoustical situation must be treated as a system comprised of three parts: source, transmission path, and listener. When the properties of the source are known, the transmission path can be modified to attenuate the sound to suit the listener's needs.

Sources. Noise sources are specified in terms of the total acoustical power radiated in each of a number (generally between 8 and 25) of contiguous frequency bands.^{1,2} A standard set of ten bands is listed in Table 1.

TABLE 1. STANDARD OCTAVE FREQUENCY BANDS

Lower and Upper Frequency Limits of Each Band (cps)	Geometric Mean Frequency of Each Band (cps)
22.1-44.2	31.5
44.2-88.5	63
88.5-177	125
177-354	250
354-707	500
707-1,414	1,000
1,414-2,828	2,000
2,828-5,655	4,000
5,655-11,310	8,000
11,310-22,620	16,000

Because of the wide range of sound powers encountered in practice, it is customary to express them in a logarithmic form. Thus we speak of the strength of a sound source in terms of *sound power level*, W , in decibels, defined by item 1 in Table 2.

TABLE 2

Decibel Scale	Abbreviation	Reference Quantity	Definition
Sound power level	PWL	$W_{\text{ref}} = 10^{-12}$ watt	$10 \log_{10} \frac{W}{W_{\text{ref}}}$ db
Sound intensity level	IL	$I_{\text{ref}} = 10^{-12}$ watt/m ² $= 10^{-14}$ watt/cm ²	$10 \log_{10} \frac{I}{I_{\text{ref}}}$ db
Sound pressure level	SPL	$p_{\text{ref}} = 0.00002$ newton/m ² $= 0.0002$ microbar	$10 \log_{10} \frac{p^2}{p_{\text{ref}}^2}$ $= 20 \log_{10} \frac{p}{p_{\text{ref}}}$ db

We note that sound power W is expressed in watts.

A listener does not experience the total sound power from a source, since it radiates in all directions, but rather the proportion that arrives at his ear. Thus we speak of *sound intensity*, I , as the sound power passing through a small area at the point of observation. The units are watts per square centimeter or per square meter. *Sound intensity level*, IL, in decibels, is defined by Item 2 in Table 2.

There is no commercially available instrument for measuring sound intensity, so it must be determined indirectly from the mean-square sound pressure, p^2 , i.e., the time average of the square of the instantaneous sound pressure in the acoustic wave. This quantity can be determined readily with a pressure microphone. The relation is given by,

$$I \approx p^2 / \rho c \text{ watts/m}^2 \quad (1)$$

where, ρ is the density of air (or other gas) in kilograms per cubic meter and c is the speed of sound in air in meters per second. *Sound pressure level*, SPL, in decibels, is defined by item 3 in Table 2.

Instruments and techniques for the measurement of sound pressure levels are widely available.² Typical measured values of sound power levels for many sources are given in references 1 to 3.

Paths. Sound may travel from a source to a receiver by many paths, some in the air (outdoors or in a room), some through walls and some along solid structures. In the latter two cases, the sound is radiated into the air from the vibrations of the surfaces.

Outdoors, the relation between the sound pressure level measured at distance r from a source and the sound power level of the source is given by,

$$\text{SPL}_\theta \approx \text{PWL} + \text{DI}_\theta - 20 \log_{10} r - 11 \text{ db} \quad (2)$$

where it is assumed that the source is near a hard-ground plane at a distance r in meters from the receiver (also near the plane) and that the source produces different sound intensities in different

directions, θ , as described by a *directivity index*, DI_θ (see reference 2). If the source radiates sound equally in all directions, then $\text{DI} = 0$. In practice, sources generally have directivity indexes in the range of 0 to 12 db in the direction of maximum radiation. At large distances, r , there will be losses in the air itself at frequencies above 1500 cps. Also, wind, temperature gradients, and air turbulence may reduce or augment the SPL_θ determined from Eq. (2).

In a room, the sound pressure level produced by a nondirective source is given by,

$$\text{SPL} = \text{PWL} + 10 \log_{10} \left(\frac{1}{4\pi r^2} + \frac{4}{R} \right) \text{ db} \quad (3)$$

where r is the distance between the receiver and the microphone and R is the *room constant* in square meters (see reference 2). Typical values of R are found in Fig. 1.

In practical cases, we are often interested in the sound pressure level produced in a room separated by a partition (wall) from the room in which the source is located. We assign a *transmission loss*, TL, in decibels to the intervening wall. Curves of transmission loss versus frequency for several different building structures are given in Fig. 2 and reference 4. The equation relating SPL to the PWL of the source is,

$$\begin{aligned} \text{SPL}_2 = \text{PWL}_1 - \text{TL} + 10 \log_{10} \left(\frac{S_w}{R_1} \right) \\ + 10 \log_{10} \left(\frac{1}{S_w} + \frac{4}{R_2} \right) \text{ db} \end{aligned} \quad (4)$$

where SPL_2 is the sound pressure level in the second room produced by a source in the first room; TL is the transmission loss; S_w is the area of the wall in square meters, and R_1 and R_2 are the room constants for the first and second rooms, respectively, in square meters. It is assumed that the sound pressure level is measured near the common wall; in the center of the room it may be 3 to 5 decibels lower.

Technique. It is apparent from Eq. (4) that the techniques for noise reduction indoors are threefold. First, make every effort to reduce the

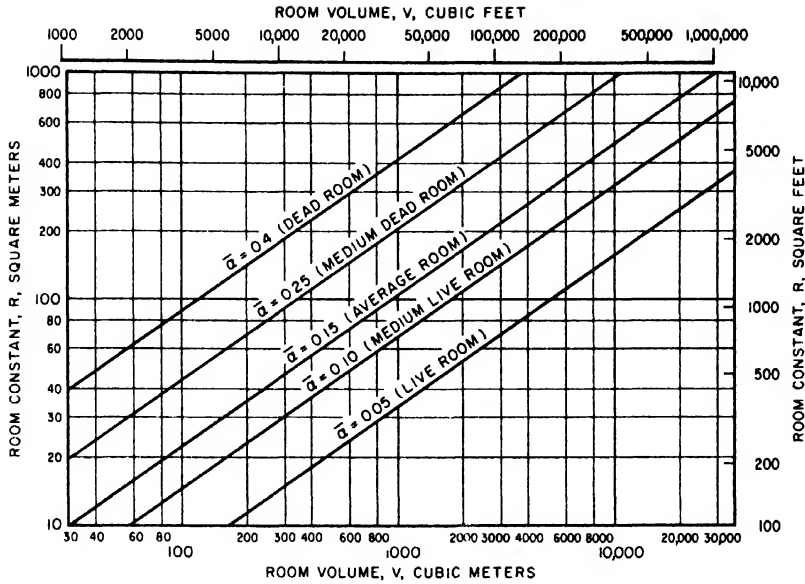


FIG. 1. Approximate value of room constant R for five categories of rooms ranging from "live" to "dead." Metric units referenced at bottom and left, English units top and right. The Greek letter $\bar{\alpha}$ indicates the percentage of the energy that is removed from a sound wave when it reflects from an "average" surface of the room. It is called the average sound absorption coefficient.

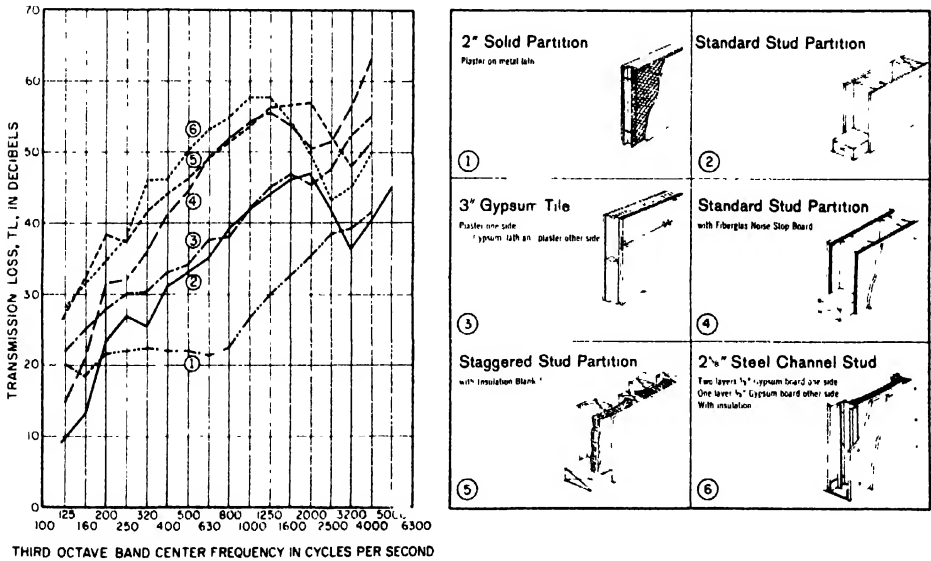


FIG. 2. Transmission loss, TL, of six typical building structures.

sound power radiated by the source, i.e., use quiet ventilating fans, quiet typewriters, quiet factory machinery, and so forth. Enclose noisy machinery in separate rooms or in enclosures. Mount vibrating machinery on resilient pads or springs. Second, provide walls with suitably high transmission losses between rooms. For example,

between adjoining apartments, walls 4 to 6 of Fig. 2 are usually satisfactory, while walls 1 to 3 are not. On the other hand, walls 2 and 3 would be satisfactory between rooms of the same apartment, while wall 1 would not. Finally, increase the room constants by adding sound-absorbing materials to either or both rooms, e.g.,

carpets and draperies, or acoustical materials on ceiling or walls or both. The sound-absorbing efficiencies of various materials are given in references 2, 3 and 5.

It is of great importance to observe that when a wall is placed between two rooms or when an enclosure is built around a noisy machine, the structure must be hermetically sealed, or, if air-flow is necessary, it must be conducted in and out of the enclosure through suitable silencers. A hole even as small in diameter as a pencil can render an otherwise satisfactory wall or enclosure inadequate acoustically.

In cases of very high noise levels where it is impractical acoustically to quiet or to isolate the machine, then ear plugs, ear cushions, or both, must be worn by personnel exposed to the noise.

Criteria for Design. Acceptable noise levels in rooms of various types in each of eight octave frequency bands are shown by Fig. 3 and Table 3.

Auditoriums for Speech. Three goals must be met in the design of auditoriums for speech. First, the ambient noise levels must be sufficiently low (see Table 3). Second, speech must be loud enough in all parts of the room so that faint syllables can be heard in the presence of normal audience noise. This second goal is achieved in small auditoriums (under about 500 seats) by proper shaping of the front part of the hall so that the speaker's voice is directed uniformly to

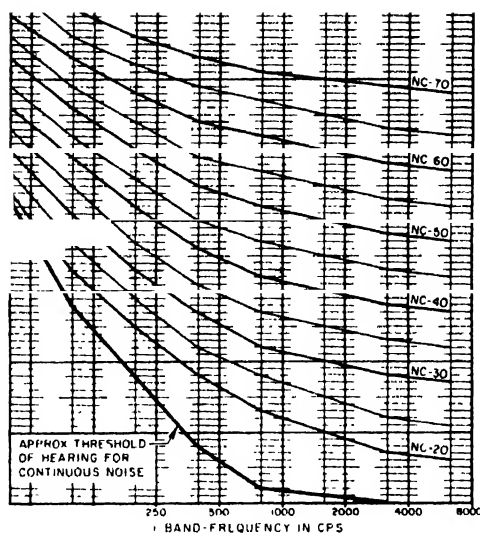


FIG. 3 Noise Criteria (NC) curves for various types of building spaces given in Table 3. Measurements are made with an octave band filter and the readings in each band should not exceed that shown on the appropriate NC curve.

TABLE 3

	Recommended NC Curve*	Communication Environment	Typical Applications
Office	30-35	"Quiet" office; satisfactory for conferences at a 15-ft table; normal voice 10-30 ft; telephone use satisfactory	Private or semi-private offices, reception rooms and small conference rooms for 20
Office	40-50	Satisfactory for conferences at a 4- to 5-ft table; telephone use occasionally slightly difficult; normal voice 3-6 ft; raised voice 6-12 ft	Large engineering and drafting rooms, etc
Office	50-55	Unsatisfactory for conferences of more than two or three people; telephone use slightly difficult; normal voice 1-2 ft; raised voice 3-6 ft	Secretarial areas (typing), accounting areas (business machines), blueprint rooms, etc.
Hall for music	15-20	All types of music	Concert halls, music rooms, broadcast studios and opera houses
Auditorium (less than 500 seats, no amplification)	20-25	Lectures with raised voice	Business, school, and hotel auditoriums
Schoolroom (no amplification)	25	Normal voice from standing or sitting position	Grammar schools, high schools, and universities
Church (no amplification)	25	Sermon from lectern	Small churches and synagogues
Apartment or hotel	25-30	Listening to TV, using telephone, and conversation	Living and sleeping areas

* NC = Noise criteria, see Fig. 3.

all parts of the hall. In large halls (over 500 seats), electronic amplification of speech is usually necessary. Third, the reverberation in the auditorium should be sufficiently low that speech is distinct. In auditoriums where there is no sound system, this requirement means that either the ceiling should have an average height of less than 30 feet above the main floor, assuming that the seats are upholstered and that there are no large floor areas without seats. If the ceiling height is over 30 feet, sound-absorbing materials will have to be added to the walls and, perhaps, the rear ceiling to control the reverberation. A satisfactory shape of a 500-seat auditorium for unamplified speech is shown in Fig. 4.

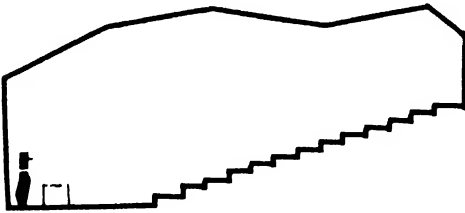


Fig. 4. Satisfactory ceiling shape for a speech auditorium with less than 500 seats.

Auditoriums for Music. There appears to be no single, ideal architectural solution for the acoustical design of a hall for music. Successful acoustics have been achieved with rectangular, fan or wedge, horseshoe, and even asymmetrical, plans. But though this is true, the many attributes of musical-architectural acoustics are so closely interrelated that if a hall is to be successful, the architect must solve all requirements simultaneously.

The music of each era of the past was composed for a different acoustical environment. Music of the Baroque period (Bach and earlier), except for organ music, was composed for small halls with relatively short reverberation times. (That is to say, a loud sound should take about 1.5 seconds to die down to inaudibility after its source is cut off abruptly. This quantity, called *mid-frequency reverberation time*, is measured with full audience present at 500 and 1000 cps and averaged.) Music of the Classical period (early Beethoven, Mozart and Haydn) was composed for larger halls with medium reverberation times (about 1.7 seconds). On the other hand, music of the Romantic period (after 1850) was, in general, composed for fairly large halls with long reverberation times (about 2.2 seconds). Today, halls must not only accommodate a musical repertoire extending over centuries, but often they must seat so large an audience that they become an entirely new type of space in which to perform music.

In the development of the design of a hall, the acoustics dictate the cubic volume and strongly influence the orientation of every sound-reflecting surface, the interior materials, and even the seating.

Concert hall and opera house design is complex,^{6,7} but some guiding principles stand out. The seating capacity should be low, below 2200 if possible. The ceiling should have an average height of 45 feet, if there are no balconies, or 55 feet with balconies, measured above the floor beneath the main floor seats. The hall should be narrow, or other means such as suspended panels should be provided for producing early sound reflections at listener's positions. Finishes for the interior should primarily be plaster. Not over 20 per cent should be wooden if the strength of the bass tone is to be preserved. Irregularities on all the surfaces should be provided to produce diffusion and blending of the sound. Above all, avoid echo, noise and tonal distortion. Finally, the orchestra enclosure should provide sectional balance in the orchestra and permit the musicians to hear each other.

Boston Symphony Hall, one of the world's best-liked concert halls, is rectangular, as shown in Fig. 5, and meets the general requirements

BOSTON, SYMPHONY HALL

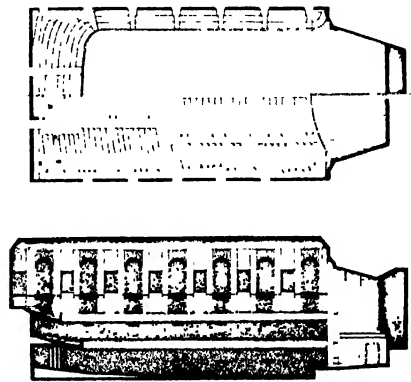


FIG. 5. Drawings of Symphony Hall, Boston, Mass.

listed above. Its mid-frequency reverberation time, with full audience, is 1.8 seconds.

LEO L. BERANEK

References

1. Peterson, A. P. G., and Gross, E. E., Jr., "Handbook of Noise Measurement," General Radio Co., West Concord, Mass., 1963.
2. Beranek, L. L., Ed., "Noise Reduction," New York, McGraw-Hill Book Co., 1960.
3. Harris, C. M., Ed., "Handbook of Noise Control," New York, McGraw-Hill Book Co., 1957.
4. "Performance Data—Architectural Acoustical Materials," Acoustical Materials Association, New York, N.Y., published annually (A.I.A. No. 39-B).
5. "Solutions to Noise Control Problems in the Construction of Houses, Apartments, Motels, and Hotels," Owens-Corning Fiberglas Corp., Toledo, Ohio, 1963 (A.I.A. No. 39-E).

6. Beranek, L. L., "Music, Acoustics, and Architecture," New York, John Wiley & Sons, Inc., 1962.
7. Furrer, W., "Room and Building Acoustics," (translated by E. R. Robinson and P. Lord), London, Butterworths, 1964.
8. Kinsler, L. E., and Frey, A. R., "Fundamentals of Acoustics," New York, John Wiley & Sons, Inc., 1950.
9. Beranek, L. L. "Acoustics," New York, McGraw-Hill Book Co., 1949.

Cross-references: ACOUSTICS; HEARING; MUSICAL SOUND; NOISE, ACOUSTICAL; PHYSICAL ACOUSTICS; RESONANCE; VIBRATION.

ASTRODYNAMICS

Astro dynamics is the study of how bodies move in space, particularly under "free fall," or the influence of gravitational forces alone, but also as influenced by nongravitational forces such as drag, thrust, electromagnetic forces, and corpuscular radiation. The science of astrodynamics is a modern application of classical celestial mechanics to space flight. Behavior of the body in a single central gravitational force field (Keplerian flight) will be treated first, and asphericity, n -bodies, and nongravitational forces will be considered as perturbations. Atmospheric aerodynamics is not included (see AERODYNAMICS).

Ballistic Trajectories. When a projectile is fired in a vacuum from the flat earth, its trajectory after leaving the gun barrel is determined by the momentum due to the initial velocity vector and the acceleration due to the gravity vector. As shown in Fig. 1, the horizontal component of

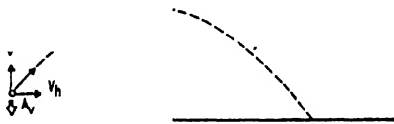


FIG. 1. Ballistic trajectory over a flat earth.

velocity remains constant until impact, while the vertical component of velocity is summed algebraically with the time integral of the acceleration due to gravity. If the vertical component of velocity is high enough, the projectile will travel so far that the flat earth assumption is not valid because the gravity vector rotates as the horizontal distance increases. In this case, the trajectory is seen to become an ellipse (Fig. 2) with the center of mass of the earth at one focus.

The range of a ballistic missile is a function of initial or burnout velocity and the elevation angle. As the burnout velocity is increased, the missile will travel farther before impact. At one critical velocity vector (25,900 ft/sec at zero-degree elevation angle), the projectile will not fall fast enough to hit the earth, but (neglecting atmospheric braking) would circle it. At this velocity, the centrifugal force exactly equals the mass attraction of the earth.

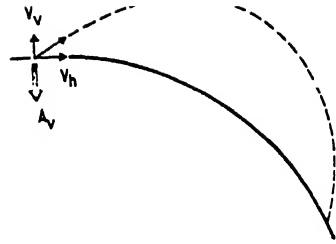


FIG. 2. Ballistic trajectory over a spherical earth.

Orbital Energy. The kinetic energy of a moving mass (m) with velocity (V) is

$$E_k = \frac{1}{2}mV^2$$

The potential energy, referenced to zero energy at an infinite distance from the attracting mass (and becoming more negative as it approaches the attracting mass), is

$$E_p = \frac{-mMG}{r}$$

where $MG = K = g_0 r_0^2$ is the gravitational constant for the attracting mass and r is the distance between the masses.

For the earth (used in satellite calculations):

$$K \approx 1.407 (10^{16}) \text{ ft}^3/\text{sec}^2$$

For the sun (used in planet and solar system travel calculations):

$$K \approx 4.679 (10^{21}) \text{ ft}^3/\text{sec}^2$$

For the moon (used in lunar orbit calculations):

$$K \approx 1.727 (10^{14}) \text{ ft}^3/\text{sec}^2$$

The total energy per unit mass, then, is

$$\frac{E}{m} = \frac{1}{2}V^2 - \frac{K}{r}$$

In a ballistic flight after power cutoff and before reentry, the total energy of a vehicle remains constant.

It was observed by early astronomers, and can be verified by integrating the equations of motion of a particle in space, that the total energy per unit mass is related to the semimajor axis a of the conic Keplerian trajectory

$$\frac{E}{m} = \frac{-2K}{a}$$

This is reduced to the famous and very useful *vis viva* energy equation:

$$V^2 - \frac{2K}{r} = \frac{-K}{a} \quad (1)$$

which is constant for any given trajectory.

This discloses the interesting fact that the semimajor axis of the orbital ellipse (a) is determined by the velocity and radius at the injection

point, or for that matter, at any instant during orbital flight. It is

$$E = -\frac{K}{a} = \frac{1}{2}V^2 - \frac{2K}{r}$$

The energy constant ($-K/a$) is also called C_a energy in the literature. It is a very important parameter:

(1) It is a measure of the total energy of an orbit (and therefore is useful in comparing orbits).

(2) It defines the size of an orbit.

(3) Given the periapsis distance (the distance at closest approach), it defines the shape of an orbit.

From the *vis viva* equation, the velocity required for circular orbit (where $a = r = r_0 + h$) at an altitude h above the earth is found to be

$$V_c = \sqrt{\frac{K}{r_0 + h}}$$

and the period is

$$P = \frac{2\pi(r_0 + h)}{V_c} = 2\pi\sqrt{\frac{(r_0 + h)^3}{K}}$$

The theoretical minimum orbit period of an earth satellite (at treetop height, neglecting air drag) is when $h = 0$ and the period is

$$P = 2\pi\sqrt{\frac{r_0}{g_0}} \approx 84.5 \text{ minutes}$$

Notice that the velocity and the period are both dependent on the gravitational attraction and the altitude. The mass of the vehicle is not a factor as long as it is negligible compared with the attracting mass.

Also, as will be shown shortly, the *vis viva* integral is a measure of the hyperbolic excess velocity—the residual velocity of a vehicle after it has left the field of the attracting mass. Notice that so far, both the orbital characteristics and the orbital energy are referenced to a particular attracting mass, and the center of mass attraction always lies on the plane of the trajectory and at the focus. Notice also that in central force field theory, the radius of the earth has no significance except in establishing the launch point and the point of impact. In addition, the fact that the launch point is rotating in space merely establishes an initial velocity vector before launching the vehicle. For this reason, it requires less energy to reach circular orbit velocity to the east than to the west. Of course, some orbital information may be given in terms of altitude which is simply the instantaneous radius minus the earth's radius.

Cotangential Orbits. If the injection velocity is greater or less than the circular velocity for that injection altitude, or if the injection angle is not zero, the orbit will not be circular. It can be

shown that the most efficient injection angle for satellite or space flight orbits is near zero.

For injection angles of zero degrees, if the velocity is less than the circular velocity, the orbit will be a subcircular ellipse with the injection point the farthest point from the center of mass attraction (focus) or apoapsis. If the injection velocity exceeds the circular velocity but is less than $V_c\sqrt{2}$, the orbit will be a hypercircular ellipse, with the injection point the nearest or periapsis point. These cotangential orbits are shown in Fig. 3. If the injection angle is not zero

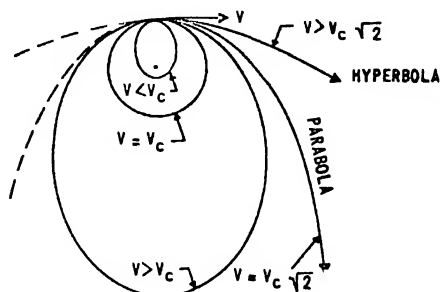


FIG. 3. Cotangential orbits.

degrees, the orbit is not a circle and the injection point is neither the apoapsis nor the periapsis.

Criterion for Escape. The kinetic energy exactly equals the potential energy when

$$\frac{1}{2}V^2 = \frac{2K}{r}$$

or

$$V = \sqrt{2K/r} = V_c\sqrt{2}$$

Therefore, if the injection velocity V_0 equals $\sqrt{2}$ times the circular velocity, the vehicle has just enough kinetic energy to overcome the potential energy of mass attraction, and it escapes from the earth or other attracting mass, theoretically reaching zero velocity at infinite distance from the center of mass attraction (assuming the vehicle is influenced by the gravity of only one body).

For $V_0 = V_c\sqrt{2}$, the trajectory is parabolic. For $V_0 > V_c\sqrt{2}$, the trajectory is hyperbolic. These trajectories are all in the family of curves called "conics"—the circle, ellipse, parabola, and hyperbola (and the trivial cases of the line and the point).

Notice that each of the closed orbits passes through its own injection point in space. The point will not generally be the same point over the earth due to the earth's rotation. Notice also that since the velocity vector and the position vector at any point on the orbit determine the entire trajectory, the orbits are dynamically reversible. Therefore, a vehicle approaching a planet with hyperbolic velocity will, if it does not impact, pass the periapsis and continue outbound on the symmetrical continuation of the approach hyperbola.

The Velocity Asymptote. The residual velocity (at an infinite distance from the attracting mass or hyperbolic excess velocity) is described by the velocity asymptote vector V_∞ which defines the injection requirements and is, therefore, a very useful parameter in describing mission guidance and energy requirements.

The square of the hyperbolic excess velocity is the C_3 energy of the orbit.

$$V_\infty^2 = C_3 = -\frac{K}{a}$$

where the semimajor axis (a) goes negative for hyperbolic flight.

For parabolic trajectories, the direction of the velocity asymptote is 180° from the perigee radius and the C_3 energy is zero. For hyperbolic trajectories, the angle between the velocity asymptote and the perigee radius is less than 180° and is a function of the C_3 energy (Fig. 4).*

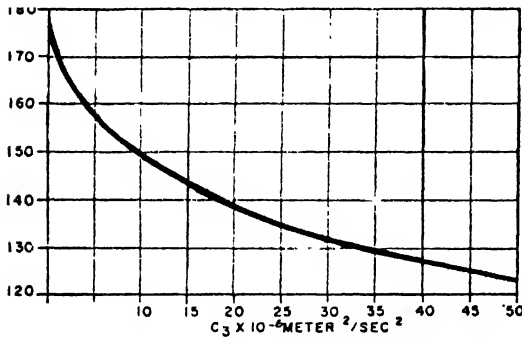
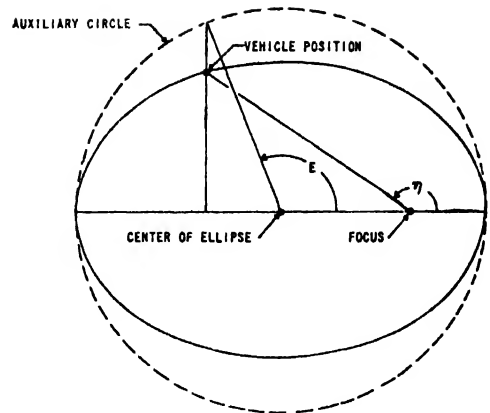


FIG. 4. Angle between outward radial (asymptote) and perigee as a function of energy.

Guidance Requirements for Space Flight. Since the laws of astrodynamics describe the flight path under gravitational forces, the function of the guidance system is to guide the vehicle during the thrust period to the required injection altitude and position and velocity vectors, and to cut off the thrust promptly.

The dynamic position of the vehicle in orbit is given by its instantaneous radius vector (r) from the attracting mass and by the true anomaly (η), which is the angle from the periapsis radius to the instantaneous position radius measured at the focus. Two other angles are very useful in relating the position of the vehicle in orbit to time: the eccentric anomaly and the mean anomaly. The eccentric anomaly (E) is the angle measured at the center of the orbit (rather than at the focus) from the periapsis radius to a point on an auxiliary circle which just contains the ellipse, the location of the point being the extension of a line perpendicular to the major axis and passing through the vehicle position. The true anomaly and the eccentric anomaly are shown in Fig. 5.

* Taken from reference 6 by permission.



η = TRUE ANOMALY
 E = ECCENTRIC ANOMALY

FIG. 5. True anomaly and eccentric anomaly.

The mean anomaly (M) is merely the time angle in radians, or the phase of the orbital period. It is given by the equation

$$M = \frac{2\pi}{P}(t - t_0)$$

where P is the orbital period, and $t - t_0$ is the time since last periapsis passage.

Kepler's Equation. The equation relating time to orbital position is given most concisely by Kepler's equation

$$M = E - e \sin E$$

Useful Orbit Relationships. A number of useful equations can be derived based on the relationship of velocity to altitude at any instant for a given attracting mass (K) and orbit size (a). It should be emphasized here that the trajectory curves are loci of the center of mass of the orbiting vehicle and are independent of vehicle attitude. The flight path direction does not describe the pointing direction of the vehicle in free flight.

Some useful orbit equations for the circle, ellipse, parabola, and hyperbola are shown below:

$$(1) \text{ Eccentricity } e = \sqrt{\frac{a^2 - b^2}{a^2}} \quad \begin{array}{l} = 0 \text{ (C)} \\ < 1 \text{ (E)} \\ = 1 \text{ (P)} \\ > 1 \text{ (H)} \end{array}$$

$$(2) \text{ Apoapsis radius } r_a = a(1 + e)$$

$$(3) \text{ Periapsis radius } r_p = a(1 - e)$$

$$(4) \text{ Semimajor axis } a = \frac{-K}{V^2 - 2K/r} = r \text{ (C)} \\ < \infty \text{ (E)} \\ = \infty \text{ (P)} \\ = -\frac{K}{V_\infty^2} \text{ (H)}$$

(5) Orbital energy constant

$$C_3 = -\frac{K}{a} = V^2 - \frac{2K}{r} < 0 \text{ (C, E)}$$

$$= 0 \text{ (P)}$$

$$> V_{\infty}^2 \text{ (H)}$$

(6) Radius

$$r = a(1 - e \cos E)$$

(7) True anomaly η

$$= \cos^{-1} \left(\frac{\cos E - e}{1 - e \cos E} \right) \text{ (E)}$$

(8) Flight path angle θ

$$= \tan^{-1} \frac{e \sin \eta}{1 + e \cos \eta}$$

(9) Velocity

$$V = \sqrt{\frac{2K}{r} - \frac{K}{a}}$$

(10) Periapsis velocity V_p

$$= \sqrt{\frac{K}{r_p} (1 + e)}$$

(11) Apoapsis velocity V_a

$$= \sqrt{\frac{K}{r_a} (1 - e)}$$

Perturbations. Perturbations are the components of acceleration of an object which are not accountable by the simple inverse-square central gravitational force field.

Examples of such perturbations to Keplerian trajectories are:

(1) The mass of the second body is not negligible.

(2) Asphericity of the principal mass (such as the equatorial bulge).

(3) Nongravitational forces—thrust, drag, electromagnetic, radiation.

The Keplerian conics are useful not only for approximations and feasibility studies but as reference orbits from which perturbations may be calculated.

The Twenty-four Hour Orbit. The circular orbit with 19,000-nautical-mile altitude is of interest because its period is 24 hours and so the satellite can be made to hover over one place on the earth.

Hohmann Transfer. Since impulsive velocity increments are more efficient than long continuous thrust to injection, the ascent to the 19,000-mile altitude is made by a low-altitude injection into an elliptical orbit with the desired apogee. At apogee, about 5½ hours flight from injection, a second injection into the circular orbit is made. This transfer is the classical Hohmann ellipse (Fig. 6).

Change of Orbital Plane. The inclination of an orbit to the equator varies with launch azimuth, but the minimum angle of inclination is the latitude of the launch site unless a change of plane is made in flight. A change of plane can be made most efficiently by a thrust applied normal to the orbital plane as the vehicle crosses the equator. If more than one orbit is involved—such as the low altitude parking orbit, the Hohmann transfer, and the high-altitude circular orbit—it can be shown that the ΔV required for change of plane is less at the highest orbit.

Station Keeping. To maintain a desired orbit, 24-hour or otherwise, an occasional small increment of velocity may be required. The velocity

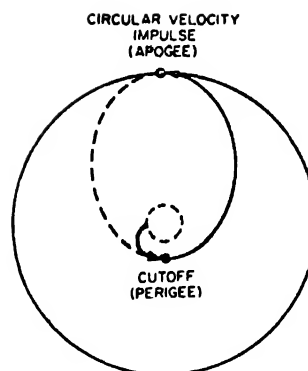


FIG. 6. Hohmann transfer ellipse.

increment to maintain a low-altitude orbit (100 to 200 miles) against the tenuous atmospheric drag is in the order of a few feet per second each week, depending on the ballistic coefficient of the vehicle.

Lunar Flight. The minimum energy flight to the moon would again be a Hohmann transfer, injecting at the perigee and reaching the moon at apogee. However, the minimum energy flight would require a 90-hour flight time. The time can be considerably reduced with relatively little increase in injection velocity. The minimum energy ellipse is very nearly parabolic ($e = 0.97$), requiring a velocity (35 860 ft/sec) very close to escape velocity, so a relatively small increase in velocity will greatly increase the semimajor axis of the ellipse beyond the moon and reduce the flight time to the moon. A flight with an injection velocity of 36 093 ft/sec would reduce the flight time to 66 hours. Instead of being at apogee, the moon would be along one side of the ellipse.

Interplanetary Flight. If planetary orbits were circular and coplanar, the minimum energy flight would be that required to provide the additional velocity to put the spacecraft into a sun-centered Hohmann transfer to Mars, or the retrograde velocity necessary for the spacecraft to fall into the smaller ellipse for a Hohmann transfer to Venus. The minimum energy flight opportunity would be when the earth at launch and the target planet at arrival have a 180° heliocentric central angle.

Synodic Periods. These opportunities occur at synodic periods due to the different orbital periods of the planets. For Mars, the synodic period is 25.6 months; Venus is 19.2 months; and Mercury is 3.8 months.

However, the planetary orbits are not circular, and they are not coplanar (Table 1); therefore, the minimum energy flights are not Hohmann transfers. The best launch dates are still very close to the synodic period (except for Mercury), but the minimum energy required is not the same at each period. Table 2 shows the minimum energy (geocentric), flight time, and planetary distance from the sun at arrival for various launch dates at the synodic period.

TABLE 1. PLANET ORBIT ECCENTRICITY AND INCLINATION TO ECLIPTIC*

Planet	Orbit Eccentricity	Inclination to Ecliptic (degrees)
Earth	0.017	0
Venus	0.007	3.39
Mars	0.093	1.85

* Taken from reference 6 by permission.

TABLE 2. CHARACTERISTICS OF MINIMUM ENERGY TRANSFERS⁵

Launch Date	Flight Time (days)	Geocentric Injection Energy ($\text{m}^2/\text{sec}^2 \cdot 10^8$)	Heliocentric Central Angle (degree)	Sun-planet Distance at Arrival (10^6 km)	Celestial Latitude of Planet on Arrival (degrees)
<i>Mars</i>					
19 Nov 64	244	0.090	174.1	231.2	0.047
5 Jan 67	202	0.091	152.2	221.9	- 0.833
2 Mar 69	178	0.088	139.3	209.7	- 1.75
24 May 71	210	0.079	156.0	216.6	0.352
30 July 73	192	0.146	141.4	234.2	1.16
<i>Venus</i>					
30 Mar 64	112	0.123	126.8	108.9	2.93
12 Nov 65	108	0.132	129.5	107.6	3.31
11 Jun 67	142	0.065	175.6	107.7	- 0.082
13 Jan 69	126	0.077	150.5	108.6	0.393
19 Aug 70	116	0.085	134.5	107.5	1.36

The energy required at each synodic period varies with the planet-to-sun distance and the distance from the ecliptic at encounter. These periodic best-launch opportunities usually last from two to five days before serious energy penalties begin.

Summary of Astrodynamical Rules. For Keplerian flight (only one force: a point gravitational attraction):

(1) The minimum velocity required for circular orbit decreases as injection altitude increases.

(2) The period of a closed orbit increases with the mean altitude of the orbit.

(3) The mean velocity of a closed orbit decreases as the mean altitude of the orbit increases.

(4) The escape velocity equals $\sqrt{2}$ times the circular orbit velocity.

(5) Orbits with injection velocity greater than circular velocity but less than escape velocity are ellipses.

(6) Orbits with injection velocity greater than escape velocity are hyperbolas.

(7) The orbital energy (hence semimajor axis, mean velocity, mean altitude, and period) remains constant in Keplerian trajectories. The energy is positive for hyperbolas, zero for parabolic escape, and becomes more negative as the semimajor axis decreases.

(8) The center of mass attraction is at the focus of the conic trajectory.

(9) Multibody trajectories can be approximated by "patched conics"—a series of Keplerian trajectories (earth-centered, sun-centered, planet-centered).

(10) Conic trajectories define the locus of the center of mass under gravitational forces. Vehicle attitude is not defined by the trajectory.

(11) The trajectory is defined by position and velocity vectors at any instant. Hence, guidance is needed only during the thrust periods.

RICHARD H. PARVIN

References

1. Baker, Robert, and Makemsen, Maude, "An Introduction to Astrodynamics," New York, Academic Press, 1960.
2. Seifert, Howard, Ed., "Space Technology," New York, John Wiley & Sons, 1959.
3. Ehricke, Krafft, "Space Flight," Vols. I and II, Princeton, N. J., D. Van Nostrand Co., 1960, 1962.
4. Nelson, Walter and Loft, Ernest, "Space Mechanics," Englewood Cliffs, N. J., Prentice-Hall, 1962.
5. Clarke, Victor C., Jr., "A Summary of the Characteristics of Ballistic Interplanetary Trajectories, 1962-1977," Jet Propulsion Laboratory Tech. Report 32-209, Pasadena, 15 January 1962.
6. Clarke, Victor C., Jr., "Design of Lunar and Interplanetary Ascent Trajectories," Jet Propulsion Laboratory Tech. Report 32-20, Pasadena, 15 March 1962.
7. Clarke, Victor C., Jr., "Constants and Related Data Used in Trajectory Calculations at the Jet Propulsion Laboratory," Jet Propulsion Laboratory Tech. Report 32-273, Pasadena, 1 May 1962.
8. Seddon, J., "Space Dynamics," *Spaceflight* (November 1963).

Cross-references: AERODYNAMICS; ASTRONAUTICS, PHYSICS OF; DYNAMICS; FLIGHT PROPULSION FUNDAMENTALS; KEPLER'S LAWS OF PLANETARY MOTION; MECHANICS; ROTATION--CIRCULAR MOTION; WORK, POWER, AND ENERGY.

ASTROMETRY

Astrometry deals with the space-time behavior of celestial bodies and therefore belongs to the classical field of astronomical studies. It is often referred to as fundamental, positional or observational astronomy.

Early astrometric investigations were directed mainly toward establishing a suitable frame of reference for the determination of the complex motions of the planets, while the studies of the positions and motions of the individual stars as well as the various stellar systems gradually developed as improved precision of observations made it possible to discover and observe these motions.

The fundamental, and perhaps most difficult, problem of astrometry is the establishment of a reference system against which the motions of the celestial bodies can be measured.

The principal planes involved in the spherical coordinate systems usually used in astrometry are the equator, defined by the rotation of the earth on its axis, and the ecliptic, defined by the revolution of the earth around the sun. The positions of both these planes vary continuously in a most complicated manner due to gravitational forces and couples between earth and the moon, the sun, and the principal planets. Such motions of the reference planes are reflected in the positions of the stars referred to them.

The motions of these planes cannot be derived entirely from theory alone, but must be deduced from observed changes in the positions of the stars which, in turn, are also in motion. This complication has forced the construction of the astronomical coordinate system to proceed by a series of successive approximations which are still in progress. Initially, the sun, and planets were observed against the "fixed stars." From these observations came the first approximations of the motions of the solar system by the laws of dynamics and of the effect of the changing orientation of the earth's axis of rotation (precession) upon the positions of the stars. Successive repetitions of the observational process have gradually improved our knowledge of these and other motions affecting the fundamental planes of the coordinate system, each improvement resulting in an increase in our knowledge of the positions and motions of the stars.

Observational programs for the improvement of the celestial coordinate system are long and tedious and must be conducted with meticulous care. They make use of highly developed instruments and observing techniques which, in combination with adopted theories of the rotation of the earth and its motion around the sun, enable the positions of the equator and equinox to be derived anew and the positions of the stars to be related to them. Each such program is an independent effort to reconstruct the celestial coordinate system. Meridian circles have generally been used for this kind of work. The results of such programs are said to be fundamental and

are usually published in the form of star catalogs.

From time to time, when sufficient fundamentally observed catalogues have accumulated, they are combined with similar earlier material to form a *Fundamental Star Catalog*. This catalog is usually regarded as the best representation that may be had of the celestial coordinate system at the time of its publication; the right ascensions and declinations of the stars in the catalog define the system for the equinox and epoch chosen for the catalog. The proper motions in combination with the adopted values of the constant of precession permit the system to be referred to equinoxes and equators at other epochs.

The latest and most precise of the fundamental catalogs is designated the *FK4* and was published by the Astronomischen Rechen-Institut, Heidelberg, Germany in 1963. The catalog contains the positions (right ascension and declination) and the changes with time (precession and proper motion) of 1535 stars. These data were compiled from nearly 200 star catalogs containing observations over a span of 110 years. Two other fundamental catalogs that have been extensively used are the *GC*, *Albany General Catalogue of 33342 Stars* and the *N30, Catalog of 5268 Stars*.

The coordinate system provided by the positions and motions of the stars in a fundamental catalog serves as a reference system for the measurement of other star positions and proper motions which must be carried out for a variety of problems originating in the study of stellar motions, in geodesy, in the determination of time, in space research and others.

With the exception of the *GC*, which contains all the stars brighter than the 7th magnitude, fundamental catalogs do not contain a complete list of all stars down to a certain magnitude as, for example, the survey catalogs do. The prototype of the survey catalogs for star positions is the *Bonner Durchmusterung* which contains the positions of 320,000 stars to a limiting magnitude of 9.5 and north of declination -2° . Although the observations for the catalog were made in the middle of the past century, the catalog and the charts made from it have been an extremely useful tool for astronomers for identification of star fields. The survey was later extended to the south celestial pole by the Bonn, Cape and Cordoba Observatories.

Positions of the fainter stars on a fundamental system are obtained by a close coordination between visual and photographic programs. The positions of a selected number of moderately bright stars (7th to 9th magnitude) are related to the fundamental system by meridian circle observations. These stars are then used as a position reference for the photographic observations of the fainter stars, thus tying them to the fundamental system.

An example of this procedure is the large astrometric project initiated toward the end of the nineteenth century and carried out by international cooperation.

The fundamental system adopted for this undertaking was embodied in the *FC (Fundamental-Catalog für die Zonen-Beobachtungen am Nördlichen Himmel)* developed by Auwers. The visual program, designated the *AGK (Astronomische Gesellschaft Katalog)*, was carried out through the collaboration of 12 northern hemisphere observatories and resulted in the determination of the positions with respect to the *FC* of 144,128 stars to the limiting magnitude of 9 and north of -2° declination. The extension of the visual work into the southern skies was gradually carried out by other observatories. The adjunct photographic program, known as the *Carte du Ciel* or the *Astrographic Catalogue* called for observations down to approximately the 11th magnitude covering the entire sky by $2^\circ \times 2^\circ$ fields. Originating in 1887, the program has only recently been completed and involved the participation of 18 different observatories. The positions in the catalogs are given in the form of rectangular coordinates as measured on the plates, but by means of auxiliary tables, these coordinates can be translated into right ascension and declination. Each field was photographed a second time with a longer exposure with a limiting magnitude of 14 to be used for the purpose of star charts.

Several other catalogs of photographically derived positions have been published. Among the catalogs of this nature may be mentioned the *AGK2 (Zweiter Katalog der Astronomischen Gesellschaft)* and the Yale and Cape photographic catalogs.

The *AGK2* was rigorously related to the fundamental system represented by the *FK3 (Dritter Fundamentalkatalog des Berliner Astronomischen Jahrbuchs)* through the use of simultaneous visual observations of about 13,000 moderately bright stars in making the plate reductions. The *AGK2* plates were taken at the Bonn and Bergedorf Observatories and covered the sky in $5^\circ \times 5^\circ$ overlapping fields from 2° to the north pole. The resulting catalog contains the positions of over 180,000 stars for the mean epoch of 1930. A second photographic series of observations of these stars, known as the *AGK3*, has been in progress at the Bergedorf Observatory since 1956. The plates will be reduced to the system of the *FK4* by use of the positions of some 21,000 reference stars observed simultaneously through an international cooperative program involving 12 meridian circles in the northern hemisphere. A comparison of the plate results at the two epochs will give rather accurate proper motions with respect to the fundamental system for the entire 180,000 stars. The majority of these stars are brighter than the 9th magnitude. A good many, however, are as faint as the 11.5 photographic magnitude.

Except for a gap between $+50^\circ$ and $+30^\circ$ declination, the Yale photographic catalogs cover the sky from $+90^\circ$ to -30° while the Cape catalogs, at present, extend this coverage to -64° declination. Cape photographic catalogs in preparation will complete the coverage to the

south pole. Both series of catalogs were taken by zones of declination by use of wide-angle cameras (from $5^\circ \times 5^\circ$ to $10^\circ \times 14^\circ$) and were reduced to a fundamental system (not always the same one) by use of contemporary meridian circle observations. The mean epochs of the positions in these catalogs range from the early 1930's to the late 1940's. The stars in these catalogs are similar in magnitude range to those in the *AGK2* and *AGK3*. Both the Yale and Cape Observatories now have plans in progress to reobserve all the stars south of the equator by photographic methods. The visual observations of about 20,000 stars needed to tie the photographic results to the fundamental system are being obtained through an international cooperative program known as the *SRS (Southern Reference Star Program)*. It is planned eventually to derive proper motions for all the stars in these catalogs through a comparison of the current photographic positions with the earlier ones.

An important source for obtaining proper motions of the fainter stars is a combination of early photographic plates with recent ones taken with the same telescope. The proper motions derived in this way are relative proper motions and require further reductions for transformation into absolute proper motions in a fundamental system. This procedure has been followed in several extensive programs aimed at solving such problems as determining the solar motion, and deriving secular parallaxes and galactic rotation.

Proper motions for tens of thousands of stars have been obtained by this method while radial velocities for a lesser number of stars have been determined from the spectroscopic application of the Doppler principle. Besides the proper motion and radial velocity, the distance of a star is needed to determine its motion in space. For stars beyond 30 parsecs from the solar system it becomes increasingly difficult to obtain all three factors involved, and often knowledge of stellar motion is either based on proper motions or radial velocities alone. However, by various statistical devices substantial information about stellar motions has been obtained.

On the basis of these studies, the sun's velocity has been determined to be about 20 km/sec towards a point in space not far from Vega, although the amount and direction of the motion varies depending upon the chosen group of stars.

As a result of the sun's motion through space, the stars show a parallactic or secular shift which can be used to determine their distances. Because of the individual motions of the stars, this method is applicable only to groups of stars with the assumption that their individual motions are random. By means of the secular parallax method, general ideas of the distances of stars up to 1000 parsecs have been obtained.

From statistical studies of proper motions and radial velocities, it was found in 1927 that the stars in our galaxy are moving in orbits not greatly inclined to the galactic equator. The observations

are consistent with the assumption that the principal force governing the motions is gravitational with the center of mass near the galactic center. The period of rotation at the sun's distance from the center is $2 \cdot 10^8$ years.

There are, at the present time, two programs in progress which will attempt to establish absolute stellar proper motions using the distant galaxies as a reference frame, the assumption being made that those objects do not show any systematic rotation with respect to the local inertial frame of rest.

A large number of proper motion studies of galactic clusters has been carried out in order to establish membership of the individual stars in the field. Because of the high internal precision required for this work, these studies have been confined primarily to long-focus telescopes, with plates taken over time intervals of 50 or more years.

Several surveys of the sky for stars with high proper motion, largely with the aim of finding absolutely faint stars, have been carried out over the past several decades. Two surveys are still in progress, one with the 48-inch Schmidt telescope at Mt. Palomar and the other with the 13-inch telescope at Lowell Observatory, the latter being by far the most extensive survey to date. It will cover 80 per cent of the sky to a limiting magnitude of 17, and in the course of its completion, upwards of 240 million star images will have been examined. In surveys of this magnitude, it is essential that telescopes of not too large focal length be used to limit the number of plates needed to cover the sky and that the "moving" stars be found by rapid scanning of the plates. These surveys have drastically increased the known number of white-dwarf, sub-dwarf, and faint red-dwarf stars, which, at the present time attract much interest among astronomers.

An important area within astrometry is the determination of the distances of individual stars. Because of the extremely small quantities to be measured, the ultimate in precision is required. The geometric method of measuring distances is based upon the surveyor's principle, the object is observed from both ends of a base line. In determining the trigonometric parallax of a star, the semimajor axis of the earth's orbit is used as the base line. Reliable individual distances have been measured in this way for several thousands of stars within 30 parsecs of the solar system. Beyond this limit, the annual parallactic effect measured against a background of more distant stars becomes so small that it cannot be measured accurately. On the plates taken with telescopes having the largest scale, the parallactic shift of stars near the limit of 30 parsecs amounts to no more than 3μ . Shifts of this order require series of photographic plates taken over several years for their measurement. This necessitates the use of telescopes with high stability in their optical systems and the application of the most refined techniques of photography and measurement. It is only since the turn of the century that sufficiently precise instrumentation and

techniques for this task have been available. The importance of stellar distance determination is realized from the fact that the distance of a star must be known before its intrinsic luminosity and its rate of energy generation can be determined.

Studies of the stars in the solar neighborhood have revealed that the majority of them are components of double and multiple systems. Since the motions of the stars within a system are governed by their mutual gravitational attraction, it is possible to determine their masses by use of Kepler's third law (see KEPLER'S LAWS OF PLANETARY MOTION), whenever their orbital motions and the parallax of the system become known. This is the only direct way masses of the stars can be determined.

Routine observations of the motions in binary systems began about 135 years ago. Originally all observations were carried out visually. Although this method continues to be used for close pairs, it has been largely replaced by a more accurate photographic method for wider pairs.

Various searches for double stars have produced some 65,000 visual binary systems, but for only a small fraction (approximately 50) of these systems are data available for determining the individual masses with an accuracy of 30 per cent or better. These masses range from about 0.08 solar mass for a star 3000 times less luminous than the sun to 6 times the sun's mass for a star 100 times more luminous than the sun. Larger masses, as high as 50 to 100 times the solar mass, are found among the very close binaries such as eclipsing and spectroscopic binaries. Although these objects cannot be resolved into individual components, their orbital motions can be determined from the periodic variation in light and radial velocity (observed Doppler shift).

Stellar masses smaller than the value of 0.08 quoted above have been discovered in recent years by intensive photographic studies of nearby single stars and components in double stars. These studies, representing the ultimate in accuracy in photographic astrometry, have revealed unseen companions of such small masses that, according to theoretical estimates, they are either stars so small that they will never burn nuclear fuel or planets of the size of Jupiter.

Aside from demands for such utilitarian purposes as navigation, geodesy and space research, astronomers themselves are making heavy demands for substantial gains in quality and quantity in astrometric observations, extended to fainter and fainter stars.

The discovery of the intrinsically faint stars in the solar neighborhood has demanded an extensive parallax program with an entirely new telescope of special design.

Positions and space velocities on a large scale of the individual stars and of stellar systems within our galaxy are essential to understand its dynamics and evolution as well as the physical properties and evolution of the individual stars which populate it.

To accomplish this, new instrumental and

analytical techniques are currently being introduced which take advantage of the latest technological developments in automation.

K. AA. STRAND

References

- Hiltner, W. A., "Astronomical Techniques," Chicago, University of Chicago Press, 1962.
 Smart, W. M., "Spherical Astronomy," Cambridge, The University Press, 1931.
 Smart, W. M., "Stellar Dynamics," Cambridge, The University Press, 1938.
 Strand, K. Aa., "Basic Astronomical Data," Chicago, University of Chicago Press, 1963.
 Trumpler, R. J., and Weaver, H. F., "Statistical Astronomy," University of California Press, 1953.

Cross-references: ASTROPHYSICS, DOPPLER EFFECT, KEPLER'S LAWS OF PLANETARY MOTION.

ASTRONAUTICS, PHYSICS OF

Future historians may record that the age of space flight marked a turning point in modern times. Its physical principles—the laws of motion of celestial bodies—marked the turning point of medieval times. From Copernicus to Kepler to Galileo to Newton, the Aristotelian myth of a man-oriented universe succumbed to the conception of a detached mechanically oriented universe, operating through laws which were a synthesis of the new knowledge gained in the formerly separate domains of terrestrial and celestial mechanics. Mankind never quite recovered from that detachment.

Weight and Weightlessness. In Newtonian mechanics, weight is understood to mean the force that an object exerts upon its support. This would depend on two factors: the strength of gravity at the object's location (things weigh less on the moon) and, as Newton called it, the quantity of matter in a body (its "mass"). At any given location, where gravity is fixed, mass can be measured relative to a standard by noting the extension of a spring to which it and the standard are successively attached. Alternatively, the unknown and standard may be hung at opposite ends of a rod and the balance point noted. However, by an entirely separate experiment, mass can also be measured by noting the resistance of the object to a fixed force applied horizontally on a frictionless table. The measured acceleration provides the required basis of comparison with the standard. Needless to say, all objects measure identical accelerations when freely falling in the vertical force of gravity. This merely means that, unlike the arbitrary force we apply horizontally in the experiment above, gravity has the property of adjusting itself in just the right amount, raising or lowering its applied force, to maintain the acceleration constant.

It was well known that objects appear to increase or decrease their weight (alter the extension of the spring) if the reference frame in which

the measurement takes place accelerates up or down. As gravity did not really change, however, most people were inclined to draw a distinction between *weight* defined as mg , where m is the mass and g is the local gravity field, and the *appearance of weight*, the force of an object on its support as measured by the spring's extension. One way to avoid the difficulty has been to speak of an *effective g*, which takes into consideration the frame's acceleration. For example, at the equator of the earth, we measure, say by timing the oscillation of a pendulum, the effective g , some 0.34 per cent less than the g produced by the mass of earth beneath our feet. If the earth were rotating with a period of an hour and a half instead of 24 hours, our centripetal acceleration at the equator would cause the effective g to vanish completely, our scales would not register, objects would be unsupported, and for all practical purposes we would be weightless.

Formally, we could state that any accelerating frame produces a local gravitational field g_{acc} that is equal and opposite to the acceleration. Thus, a rotating frame generates a centrifugal g_{acc} opposing the centripetal acceleration. We have at any point

$$g_{eff} = g + g_{acc} \quad (1)$$

where g is the field produced by matter alone (e.g., the earth). By identical reasoning, an object in orbit, whether falling freely in a curved or in a straight path, will carry a reference frame in which g_{eff} is zero, for its acceleration will always exactly equal the local g by the definition of the phrase, "freely falling."

This concept was placed on a firm footing by Einstein who maintained that Eq. (1) is reasonable not only in mechanics but in all areas of physics including electromagnetic phenomena. We arrive at the inevitable conclusion that we cannot distinguish by any physical experiment between an apparent g accountable to an accelerating frame and a "real" g derived from a local accumulation of mass. This central postulate of the General Theory of Relativity also unified the two separate conceptions of mass. An object resting on a platform that is accelerating toward it will resist the acceleration in an amount depending on its inertia. It presses against the platform with a force equal to that it would have if placed at rest on the surface of a planet with local field equal and opposite to the acceleration of the frame.

General Principles of Central Force Motion. The gravitational force between point masses is inverse square, written

$$mg = -\frac{\gamma m' m}{r^2} \hat{r} \quad (2)$$

where the center of coordinates from which the unit vector \hat{r} is described lies in m' , one of the masses. Thus, the force on m is directed $-\hat{r}$, toward m' and is proportional to $1/r^2$ with γ the constant of proportionality. The quantity g is the force on m divided by m , (or normalized force) for which the name "gravitational field of m' "

is reserved. Of course, if m were in the field of a collection of mass points, or even in a continuous distribution of mass, the summated or integrated g at the location of m would no longer be an inverse square function with respect to any coordinate center. However, in one special case, the inverse square functional form would be preserved: if the source mass were symmetrically distributed about the coordinate center. This would be the case if the source were a spherical shell or solid sphere, of density constant or a function only of r . The sun and earth can be regarded, at least to a first approximation, as sources of inverse square gravitational fields.

There are some important general statements we can make about the motion of an object placed with arbitrary position and velocity in a centrally directed force field, i.e., a field such as the one described, which depends only on distance from a central point (regardless of whether or not the dependence is inverse square). As the force has only a radial and no angular components, it cannot exert a torque about an axis through the center. This means that the initial angular momentum is conserved. Now angular momentum is a vector quantity and therefore is conserved both in direction and magnitude. It is defined $\vec{h} = \vec{r} \times \vec{p}$, where \vec{r} is the position vector to the mass of momentum \vec{p} . The direction of the angular momentum vector is thus perpendicular to the plane containing \vec{r} and \vec{p} . As this direction is permanent, so also must be the plane. The planar motion of the object can be expressed in polar coordinates, so that, by writing $\vec{r} = r\hat{r}$ and $\vec{p} = m(\dot{r}\hat{r} + r\dot{\phi}\hat{\phi})$, we find the specific angular momentum (angular momentum per unit mass) called h , to be

$$h = r^2\dot{\phi} \quad (3)$$

This too then must be a constant of the motion.

Consider now the rate at which area is swept out by the radius vector, dS/dt . We recall from analytic geometry that $dS = \frac{1}{2}r^2d\phi$. Thus

$$\frac{dS}{dt} = \frac{h}{2} \quad (4)$$

so that this is a constant of the motion as well. On integration, we conclude that the size of a sector that is swept out is proportional to the time required to sweep it out. In the case of a closed orbit, the total area S would then be related to the specific angular momentum as

$$S = \frac{hT}{2} \quad (5)$$

This sector area-time relationship is Kepler's second law of planetary motion which was induced from Tycho Brahe's observation of Mars without prior knowledge of gravity and its central character.

The Laws of Kepler. Kepler stated two other laws of planetary motion: The orbits of all the planets about the sun are ellipses (a radical departure from the circles of Copernicus), and

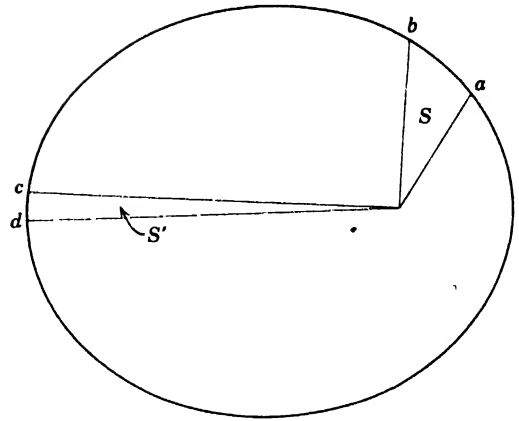


FIG. 1. Kepler's second law. The sector area S swept out is proportional to the time required for the planet to move from a to b . Thus, if $t_{cd} = t_{ab}$, then $S' = S$ (from Berman A. I., "The Physical Principles of Astronautics," New York, John Wiley & Sons, 1961).

the squares of their period are proportional to the cubes of their mean distance from the sun, this mean being the semimajor axis of their ellipses. The third law pertained to the one characteristic common to all the planets: the sun. Taken together, the three laws led Newton to the concept of gravitational force and its inverse-square form.

By applying Newton's law of motion $\mathbf{F} = m\mathbf{a}$, a relationship between \mathbf{a} , the second derivative of the position vector, expressed in polar form, and \mathbf{F}/m or g , as given by Eq. (2), leads to the familiar conic solution for the trajectory of an object in an inverse square field,

$$\frac{1}{r} = \frac{\gamma m'}{h^2} + A \cos(\phi - \phi_0) \quad (6)$$

where A and ϕ_0 are constants. A rotation of axis will eliminate ϕ_0 , thereby aligning the coordinate axis with the conic's major axis. Also, by expressing the general conic, an ellipse or hyperbola, in terms of the usual parameters of semimajor axis a and eccentricity ϵ , we can relate the geometric parameters to the gravitational-dynamical constants, viz:

$$h = [\gamma m' a (1 - \epsilon^2)]^{1/2} \quad (7)$$

and

$$\frac{1}{r} = \frac{\gamma m'}{h^2} (1 + \epsilon \cos \phi) \quad (8)$$

Note that by substituting Eq. (7) into Eq. (5) and expressing the area of an ellipse as $S = \pi a^2 (1 - \epsilon^2)^{1/2}$ we arrive at Kepler's third law,

$$T = \frac{2\pi}{(\gamma m')^{1/2}} a^{3/2} \quad (9)$$

The energy of the orbiting object can be calculated with ease by evaluating it at an extremal point, say the nearest point to the gravitational

source, called pericenter or perifocus. As the energy is constant, it is immaterial where the calculation is made. Here the velocity has only an angular component so that the kinetic energy for a unit orbiting mass is $\frac{1}{2}v^2 = \frac{1}{2}r^2\dot{\phi}^2$. The potential energy at pericenter is $-\frac{\gamma m'}{r_{pe}}$ where r_{pe} is the distance of the unit mass from m' , the focal point. Here $\phi = 0$ so that by Eq. (8),

$$\frac{1}{r_{pe}} = \frac{\gamma m'}{h^2} (1 + \epsilon) \quad (10)$$

On substituting Eq. (7), we find the total kinetic and potential energy to be

$$E = -\frac{\gamma m'}{2a} \quad (11)$$

Our conclusion: All objects in orbit with the same major axes have identical periods and identical energies per unit mass. Knowledge of E is invaluable in determining an object's speed when its distance from the source is known, and *vice versa*.

In the event that the orbiting object's mass is not negligibly small compared with that of the gravitational source, one must take note that the combined center of mass, from which the acceleration is described, no longer may be assumed to lie in the center of the gravitational source. This complicates our equations somewhat, for the accelerating force still is expressed relative to the center of the source (if spherical). The adjustment that results, when center of mass coordinates are transformed to relative coordinates in the expression for acceleration, requires our equations to take the form $\gamma(m' + m)$ wherever formerly $\gamma m'$ appeared.

Disturbances in the Central Field. The earth, of course, is spherical only to a first approximation. More accurately, it is an ellipsoid of revolution about a minor axis—an oblate spheroid.

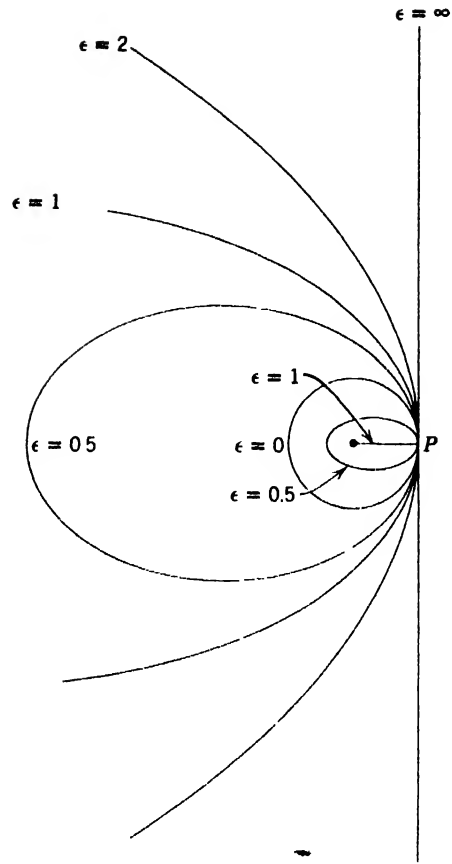


FIG. 2. Orbits of differing eccentricities and major axes which pass through a common point. Higher speeds correspond to higher energies and longer major axes (from Berman, A. I., "The Physical Principles of Astronautics," New York, John Wiley & Sons, 1961).

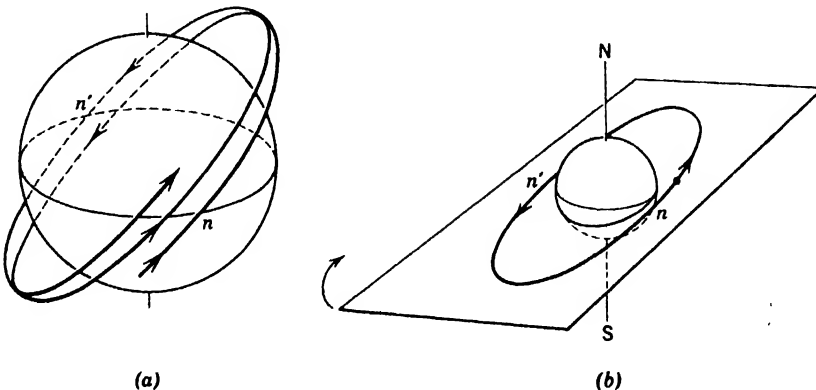


FIG. 3. The orbit of an earth satellite. The earth's equatorial bulge causes retrograde motion of the points of intersection n and n' of the orbit and equatorial plane. This can alternatively be interpreted as a retrograde motion, about the north-south axis, of the plane containing the closed orbit. The plane moves in the direction shown by the arrow in (b), maintaining a constant angle with the axis (from Berman, A. I., "The Physical Principles of Astronautics," New York, John Wiley & Sons, 1961).

Still more accurately, it appears to be slightly pear-shaped and, in addition, its figure is distorted by continuous local variations. The spheroidal figure, nevertheless, accounts for nearly all the anomalous effects of satellite orbits. For one thing, the gravitational force on the satellite is no longer centrally directed; the excessive mass in the equatorial plane produces a force on the satellite directed out of its orbital plane. The resultant torque causes the direction of the angular momentum vector to change; i.e., the plane containing the satellite's ellipse turns. The plane turns continuously about the polar axis maintaining its angle with the axis and with the equatorial plane constant. The turning rate is greatest for low orbits and small angles of inclination with the equator. For polar satellites, the plane remains fixed. A separate effect of this equatorial bulge perturbative force is the slow turning of the ellipse's major axis *within* the orbital plane. This effect vanishes at an inclination of 63.4° ; the major axis turns backward at inclinations above this angle and forward below.

Rocket Propulsion. A rocket operates by the simple principle that if a small part of its total mass is ejected at high speed, the remaining mass will receive an impulse driving it in the opposite direction at a moderate speed. As δm_e , the propellant, leaves at speed v_e with respect to the rocket, the remaining rocket mass m receives a boost in speed δv such that

$$\delta m_e v_e = m \delta v \quad (12)$$

If additional equal propellant mass is ejected at the same speed, the boost in rocket speed is slightly greater than before as the rocket mass has been slightly depleted by the prior ejection. Indeed, if the residual rocket mass eventually were minuscule, its boost in speed could reach an enormous value. The integrated effect of these nonlinear boosts is found as

$$v_t - v_0 = v_e \log_e \frac{m_0}{m_t} \quad (13)$$

where v_0 and m_0 are the rocket speed and mass at some arbitrary initial time and v_t and m_t are the same quantities at some time t later.

From these simple considerations, it is apparent that the highest rocket velocities are attained if we could increase the propellant speed as well as the mass ratio m_0/m_t . The mass ratio can be maximized by obvious methods such as choosing a high-density propellant which cuts the tankage requirement or avoiding unnecessarily complicated apparatus for ejecting propellant at high speed. A nuclear rocket, for example, may perform well in its ability to eject propellant an order of magnitude higher in velocity than conventional chemical rockets; nevertheless, the penalty required in reactor weight and shielding severely limits its effectiveness.

Specific impulse is one performance characteristic which applies to the propellant's ability to be ejected at high speed regardless of the weight penalty required to do this. It is the impulse produced per mass of propellant ejected, or $m \delta v / \delta m_e$.

or, by Eq. (12), simply v_e . In engineering usage, it is impulse per *weight* of propellant ejected, or v_e/g_e where g_e is the acceleration of gravity at the earth's surface. Its units are seconds, and it can be interpreted as the thrust produced by a rocket per weight of propellant ejected per second. By itself, thrust is of little importance unless it is sustained for a significant time by a large backup of propellant tankage. It is here that the mass ratio term in Eq. (13) would play an important role in any evaluation of a rocket's true performance.

Transfer Orbits. If one wishes to leave one orbit and enter another by rocket, an optimum path is generally chosen to minimize the total propellant required. Nevertheless, this should not be done at the expense of unduly long flight times, complicated guidance equipment, or high acceleration stresses. These would require unprofitable weight expenditures which would offset the frugality in propellant tankage.

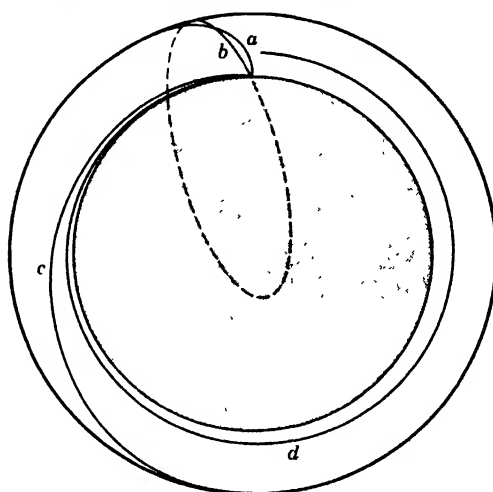


FIG. 4. Four launch trajectories into a satellite orbit about a planet. (a) If that planet has an atmosphere, the rocket may ascend in a "synergic" trajectory from the planetary surface to the final orbit, i.e., it cuts through the denser portions in an initially vertical path and gradually bends over into a horizontal path during burnout. (b) If there is no atmosphere it may ascend from the ground in a ballistic ellipse. This same ascent path may be chosen if the departure is from a parking orbit or "space platform" close to ground level. A far better choice would be (c) the Hohmann ellipse, with pericenter at the planet's surface and apocenter at the satellite orbit. Burnout time is assumed short in both this and the ballistic case. (d) A vehicle such as an ion rocket, which can sustain a microthrust for a very long time, cannot be launched from the ground but only from a parking orbit. It will spiral out to the desired altitude with few or many turns about the planet, depending on the magnitude of the thrust relative to that of the gravitational force (from Berman, A. I., "The Physical Principles of Astronautics," New York, John Wiley & Sons, 1961).

Let us examine a simple but recurring example of a transfer problem, that of leaving a space platform in one circular orbit and entering another larger one concentric with the first. If the transfer path were radial or near radial (a so-called ballistic orbit) then one would have to launch at a large angle to the direction of motion of the platform, accomplished only by a velocity component opposed to the platform's motion. On reaching the outer platform, a soft landing can be made only by a substantial rocket velocity boost tangent to the orbit. Clearly, the total propellant expenditure would be far greater than one alternative of launching the rocket in the direction of motion of the first platform with just sufficient speed to reach the outer circle, timed so that the outer platform will meet the spacecraft. The transfer orbit will be an ellipse cotangent with both circles. The outer platform will be moving much faster of course at the contact point as the major axis of its orbit is much greater [see Eq. (11)], but the difference in speed is not nearly as pronounced as for the ballistic transfer case. A differential speed increment at contact completes the maneuver.

The return trip, from an outer to inner circle, is made by following the second half of this cotangent ellipse, named the Hohmann transfer orbit after the German engineer who discovered its optimal property with regard to propellant expenditure. In the return case, the spacecraft is launched in opposition to the outer platform's motion. This removes kinetic energy and forces the spacecraft to fall in closer to the attractive center in order to make cotangent contact with the inner circle. The total propellant expenditure from the outer to the inner platform is the same as for the original journey.

An interesting question arises if one wishes to leave a platform for an outer orbit when it initially is in an elliptical orbit rather than a circle. Should we depart from apocenter where we are furthest from the gravitational source and closest to our destination? Or should we depart instead from some other point in the ellipse? Paradoxically, our best launch point is at pericenter, for here the largest possible amount of energy will be transferred to the spacecraft for a given expenditure of propellant. A given thrust applied for a given time interval will do more work on the spacecraft when it is moving fast, as at pericenter, for it covers a greater distance during the interval. This advantage offsets the undesirability of being at a lower potential energy point at pericenter.

Powered Trajectories. In the usual operation of a solid- or liquid-propelled rocket, the propellant is depleted in a time negligibly small compared with the total flight time. The trajectory analysis may generally be considered as that of a free orbit subject to burnout initial conditions as in the discussion above. If, however, the propellant ejection is sustained over long periods, as in an ion-propelled rocket, the trajectory analysis is necessarily complicated, for, in addition to the varying gravitational force, the vehicle, of slowly

diminishing mass, is subject to a thrust which may be changing both in direction and magnitude. Even one of the simplest thrust programs, a constant thrust in the direction of motion, requires an electronic computer analysis in order to obtain the position and velocity at future times (see ELECTRIC PROPULSION).

The continuous-thrust trajectory is a spiral with many advantages over the orbital ellipses. First, the lower sustained thrust precludes the high acceleration stresses associated with rapid-burning chemical rockets. Much of the structural weight usually needed to withstand these stresses can be replaced by propellant. Also, flights to the extremities of a gravitational region may take a shorter time in a spiral trajectory. In a long Hohmann ellipse, for example, most of the journey is made at very low speed. In a powered spiral, on the other hand, the spacecraft could be made to move fast, for the thrust, though small, is integrated over many months.

The spiral concept is ideal for rockets where very high ejection velocities are feasible by using electromagnetic or electrostatic particle accelerators, but only at the expense of a low propellant flow rate and relatively heavy power-generating equipment. However, the propellant reserve, and thrust, could then last the required long time. Such an ion rocket with its very low thrust-to-weight ratio could hardly be expected to take off from the ground, and could only take off from an orbital platform. In the vacuum of space, the ion beam meets its ideal environment.

ARTHUR I. BERMAN

References

- Berman, Arthur I., "The Physical Principles of Astronautics," New York, John Wiley and Sons, 1961.
 Moulton, Forest Ray, "An Introduction to Celestial Mechanics," New York, The Macmillan Co., 1914.
 Danby, J. M. A., "Fundamentals of Celestial Mechanics," New York, The Macmillan Co., 1962.
 Sterne, Theodore E., "An Introduction to Celestial Mechanics," New York, Interscience Publishers, 1960.

Cross-references: AERODYNAMICS; ASTRODYNAMICS; DYNAMICS; ELECTRIC PROPULSION; FLIGHT PROPULSION FUNDAMENTALS; GRAVITATION; INERTIAL GUIDANCE; KEPLER'S LAWS OF PLANETARY MOTION; MASS AND INERTIA; ROTATION—CIRCULAR MOTION; WORK, POWER, AND ENERGY.

ASTROPHYSICS (See also ASTROMETRY, COSMOLOGY, SOLAR PHYSICS)

Starting with the advent of photography and the study of stellar spectra in the second half of the nineteenth century, astrophysics now includes optical and radio observations of stars, clusters, interstellar material, galaxies and clusters of galaxies, and their interpretation. Radiation from these external sources provides information on

the direction of the source, its velocity, composition, temperature and other physical conditions, including magnetic fields, density, degree of ionization, and turbulence. The term "astrophysics" is generally understood to include all these aspects except the measurement of direction (positions of stars in the sky and changes due to parallax and proper motion), and the orbits of planets, asteroids and comets (celestial mechanics). Because of its proximity, the sun can be studied in more detail than other stars; its structure and its influence on the nearby planets and comets are the concern of solar physics and are closely related to geophysics and stellar astrophysics. Study of the motions of stars in pairs, groups, clusters, associations, and galaxies is the overlap of celestial mechanics with astrophysics, and the study of the distribution and patterns of motion of the distant galaxies is the overlap with "COSMOLOGY."

Early studies of stellar spectra revealed differences due primarily to surface temperature and described by the sequence of spectral types ranging from "O" (30 000°K or more) through "B," "A," "F," "G," and "K" to "M" (2000 to 3000°K). The type of a spectrum is set by relative intensities of lines and bands due to ions, atoms, and molecules. The earth's atmosphere limits the wavelength region observable from terrestrial observatories since ozone and other constituents are opaque at wavelengths $\lambda < 3000\text{\AA}$, and water vapor bands block much of the region $1.2\mu < \lambda$

8μ . The ionized layers block radio waves longer than 20 meters. Nevertheless, thousands of spectrum lines, mostly in absorption, have been identified in the range $3000 < \lambda < 12000\text{\AA}$, and the pattern of lines within one spectral type has been found to vary with a second parameter, the "luminosity class" designated by roman numerals "I" (highly luminous "super giants") through "II," "III," "IV" to "V" ("dwarfs" of relatively low luminosity).

The continuum between spectral lines has an intensity distribution with wavelength that roughly matches Planck's theoretical distribution for a blackbody, $B(\lambda) d\lambda = 2hc^2\lambda^{-5}(e^{hc/\lambda kT} - 1)^{-1}$ where T is the temperature that accounts approximately for the ionization and excitation of atoms producing the star's line spectrum, or the dissociation of molecules producing bands—that is, for the spectral type. In so far as the lines can be ignored, the color of a star is approximately that of a blackbody of temperature T and the total ("bolometric") luminosity is given by $L_b = 4\pi R^2 \sigma T^4$, where R is the radius of the star and σ is Stefan's constant. The "apparent brightness" of a star (observed optical flux) depends upon its distance, D , and its luminosity in the wavelength region observed—approximately $4000 < \lambda < 6500\text{\AA}$ for visual observations and $3700 < \lambda < 5000\text{\AA}$ for photographic observations through glass optics. Recently the introduction of photoelectric equipment has allowed more accurate measurements in smaller wavelength regions. Standard measures are designated U (ultraviolet), B (blue), V (visual), I (infrared), etc., and are

usually expressed in magnitudes, an inverse, logarithmic scale. By successive approximations in such measurements of many stars, it has been possible to correct for the effects of interstellar absorption and limited wavelength range, as well as for the inverse-square law ($1/D^2$), to obtain total luminosities and colors. The luminosities are often expressed in "suns", that is, multiples of the sun's luminosity (about 4×10^{33} ergs/sec). Analogous measurements at radio frequencies are expressed as the flux in watts/square meter/cycle/second. Spectrophotometric measurements from rockets and artificial satellites above the earth's atmosphere have been made in the far ultraviolet, and it is to be expected that the intensity distribution, $I(\lambda)$, will soon be observed for all wavelengths.

The *masses* of stars are determined from motions of double stars, ranging from widely separated visual binaries whose relative motions can be photographed, to close spectroscopic and eclipsing binaries with orbits calculated from variations in radial velocity or observed Doppler shift. Among some 50 pairs, masses of individual stars are found from 0.08 to 20 solar masses, and there is less definite evidence of others as low as 0.03 and as high as 50 or 100. (One solar mass is 2×10^{33} gm.) Over most of this range the luminosity L is proportional to M^3 .

Astrophysical theory has achieved considerable success in explaining the spectra of stars by theoretical models of the atmospheres, involving the surface temperature, surface gravity, abundances of chemical elements, turbulence, rotation, and magnetic fields. The strengths of absorption lines are found to fit a "curve of growth," the relation between measured line strengths and the strengths predicted by quantum theory for unit abundance of the one ion, atom, or molecule involved. The theory of stellar interiors further relates mass M , luminosity L , and radius R with chemical abundances, the opacity of the material, and the generation of energy by nuclear reactions. More spectacularly, it has explained stellar evolution in terms of changes due to nuclear reactions.

The theoretical models of *stellar interiors* are based on stability and two modes of transferring energy outward to the surface: radiative transfer and convective transfer. Radiative transfer, by repeated emission, absorption, and emission of light, implies a temperature gradient dependent on the opacity and on the flow of radiative energy, or L . After calculating the opacity of gaseous stellar material (about 60 per cent hydrogen, 35 per cent helium, and 5 per cent heavier elements, by weight) at various temperatures and densities, the astrophysicist can compute the temperature, density and pressure in shells at various depths inside a star, starting with a definite radius and surface temperature, and adding up the shell masses to get the total mass. At some level, the temperature and density are sufficiently high for nuclear reactions to take place, the simplest being conversion of hydrogen to helium generating 7×10^{18} ergs of energy per gram. The rate of energy generation over the

whole inner core must match L . Calculations for convective transfer follow a similar pattern but depend upon matching the adiabatic gas law for over-all stability. A combined model may have a convective core surrounded by a radiative shell and that surrounded by an outer convective shell.

The successful fitting of nuclear energy generation into such gas-sphere models of stars by 1940 led to the idea of *stellar evolution*, the conversion of hydrogen to helium in its core causing a star to age. Direct evidence of this aging was first obtained from clusters of stars. The stars in one cluster, relatively close together in space, are assumed to have been formed at the same time, and the pattern of stellar characteristics differs from one cluster to another in a systematic way. The pattern is easily recognized on a Hertzsprung-Russell ("H-R") diagram of $\log L$ vs spectral type (or color) on which the vast majority of stars appear near a diagonal line, the "main sequence."

Several other classes of stars can be distinguished by location on the H-R diagram ("red giants," "white dwarfs," etc.) and theories of stellar evolution account for a change in location along an "evolutionary track" for any one star. The rate of such change will vary in general; for instance, the large-mass, high- L blue stars are expected to exhaust their hydrogen in a few million years, whereas yellow and red dwarfs remain for billions of years on the main sequence. H-R diagrams for clusters confirm the aging (also the theoretical star models and the assumption of cluster origin) and provide evidence of the age of each cluster.

A large part of astrophysical research is devoted to filling in the details of stellar evolution and the variety of nuclear reactions involved. For example, after its hydrogen is exhausted, the core of a giant star contracts and heats up to a billion degrees; then helium combines to form carbon, providing a new intense source of energy, and gas is probably blown off the star. Later, a small white dwarf remains.

The *formation of a star* starts with gravitational contraction of a large cloud ("nebula") of interstellar gas and dust, including matter ejected from previous generations of giant stars. The recycling of material back and forth from nebulae to stars involves changes in composition—the abundances of helium and heavy elements increasing with time. Since 1954, astrophysicists have therefore been concerned with *nucleogenesis*, the creation of the chemical elements in stars (or in an early stage of the evolving universe). Differences in composition of stars in various locations are now interpreted as evidence of past star making.

The *formation of the solar system* (sun, planets, asteroids, meteors, and comets) is one case of star formation studied in great detail by astrophysicists, geologists and chemists. Radioactive dating of minerals in the earth and meteorites places this event about four billion years ago when a slowly rotating nebula contracted, forming earth and planets, but losing a good deal of its mass in the process. Fractionation of chemical elements and compounds during the condensation

is linked with astrophysical interpretation of chemical analyses of meteorites and terrestrial minerals.

Interstellar material in the form of bright nebulae has been known since telescopes were first used. During the first two decades of this century, evidence was collected showing less obvious clouds of dust and interstellar gas in the plane of the Milky Way, based on the obscuring and color effects of dust and the spectral absorption lines of gas (primarily sodium and ionized calcium). In 1945 the polarization of starlight caused by the interstellar dust was discovered, and in 1950 the radio telescope added the emission by interstellar atomic hydrogen at 21-cm wavelength. This interstellar medium is now known to extend in a thin, flat slab centered in the Milky Way. When highly luminous blue stars are in or near it, the gas is ionized by ultraviolet radiation, and the resulting electrons produce emission lines of hydrogen, oxygen, helium and other elements by recombination or by electron excitation. In addition to such "H II regions," the dimensions of which depend on the temperature and luminosity of the exciting star and the density of the medium, astrophysicists have studied more complex nebulae in which the material density varies from one place to another. The interstellar medium is often denser near young clusters or individual blue stars, as expected from the theory of star formation.

The whole *Milky Way system* of stars, nebulae and interstellar gas and dust is assumed to be in dynamic equilibrium; that is, the mass distribution can be calculated from individual motions under the gravitational attraction of the whole galaxy, another important part of modern astrophysics. Since 1920 the dimensions of the galaxy have been determined from distances of the large bright globular clusters and from the distances of nearer stars. The resulting model, a flat disk with a high-density nucleus (total mass about 10^{11} suns) also fits the average motions of stars within a few thousand light years' distance from the sun, and the radial motions of cold atomic hydrogen out to 50,000 light-years determined by Doppler shifts in the 21-cm radio emission line. The stellar motions are derived from statistics of Doppler shifts and changes in direction, allowing for random individual motions differing from the general circulation. Most of the stars in the Milky Way share in a circular velocity, v_r , around the center of the galaxy, and the mass distribution is inferred from the observed decrease of v_r with distance from the center. The globular clusters and many other stars appear to move in orbits at high inclination to the Milky Way plane, forming a "halo" around the center having little or no angular momentum. It thus appears that there are two populations in the galaxy: "Population I" stars, nebulae, gas and dust in the outer parts of a thin rotating disk, and "Population II" stars in the nonrotating halo, probably formed at an earlier time.

The many *other galaxies* well outside our own are found to include some (classed as "spirals")

very similar to our Milky Way galaxy in structure and internal motions. Others are strikingly different (classed as "ellipticals"), probably due to different conditions of formation. All the techniques of astrophysics are being applied to the study of these objects: measurement of their sizes, luminosities, colors and masses, their proportion of interstellar material, the formation and evolution of their stars, etc. These physical characteristics are fairly well correlated with morphological type (spiral or elliptical), but the masses are found to be larger than expected for the measured luminosities. Moreover, some show signs of violent explosions, possibly due to extremely high densities reached in a collapsing mass of millions of stars.

Galaxies are observed in increasing numbers at larger and larger distances, roughly in uniform distribution but with marked clustering. There is some evidence that compact clusters contain a preponderance of ellipticals and that the small groups and pairs of galaxies tend to contain galaxies of only one morphological type.

The spectra of distant galaxies show large red shifts which, if interpreted as Doppler shifts, indicate a recessional velocity proportional to distance. Here astrophysics leads into COSMOLOGY and to the mathematical models of the universe based on general RELATIVITY. Observations have yet to confirm one or another of several different cosmological models, including one based on an assumed "continuous creation" of matter.

THORNTON PAGE

References

- Abel, G. O., "Exploring the Universe," New York, Holt, Rinehart and Winston, 1964.
 Struve, Otto, and Zeebergs, "Astronomy of the 20th Century," New York, Macmillan, 1962.
 Page, Thornton, Ed., "Stars and Galaxies," Englewood Cliffs, N. J., Prentice-Hall, 1962.
 Aller, L. H., "Astrophysics," Vols I and II, Ronald, 1953, 1963.
 Hiltner, W. A., "Astronomical Techniques," Univ. of Chicago, 1962.
 Allen, C. W., "Astrophysical Quantities," Athlone Press, 1963.
 Burbidge, Burbidge, Fowler, and Hoyle, "Synthesis of the Elements in Stars," *Rev. Mod. Phys.* **29**, 547 (1957).
 Liller, William, "Space Astrophysics," New York, McGraw-Hill Book Co., 1961.

Cross-references: ASTROMETRY, COSMOLOGY, DOPPLER EFFECT, RELATIVITY, SOLAR ENERGY SOURCES, SOLAR PHYSICS

ATOM. See ATOMIC PHYSICS.

ATOMIC AND MOLECULAR BEAMS

This field of research utilizes a collision-free stream of neutral atoms or molecules as they traverse a vacuum chamber. With 10^{-6} mm Hg pressure in a vacuum chamber, air molecules at

room temperature travel on the average about 300 meters between collisions and move with an average speed of about 500 m/sec. Between collisions these molecules are essentially "free" and unperturbed by molecules of the residual gas or by atoms in the walls of the apparatus. The mathematical description of such isolated systems is much less complicated than for denser gases, liquids, or solids containing interacting particles.

Since 1911, when Dunoyer proved that a stream of neutral atoms would remain collimated in a vacuum, atomic- and molecular-beam research has become one of the most versatile, precise, and sensitive techniques for studying the properties of isolated atomic systems and interactions between such systems. Numerous fundamental discoveries in beam research have contributed to the present understanding of physical laws. The earliest experiments (1920) sought the molecular velocity distribution, which is important in the kinetic theory of gases. Atomic diameters (cross sections), van der Waals' interaction potentials, and polymer vapor composition were obtained after later refinements of technique. The Stern-Gerlach experiment (1924) demonstrated the validity of space quantization of angular momentum and established the electron spin as 1:2. This historic work placed quantum mechanics on a firmer foundation and initiated beam investigations of atomic and nuclear electromagnetic properties. Although many low-precision results appeared in subsequent years, high-precision spectroscopy began in 1937 with the introduction of the magnetic-resonance method by Rabi. In this method, transitions between quantum states separated by an energy $h\nu$ are induced by a radio-frequency field of frequency ν (h is Planck's constant). From the Heisenberg uncertainty principle, the width of the rf resonance is small, owing to the long lifetimes of the beam quantum states and to the ability to irradiate the beam with radio frequency for as long as a few milliseconds along its path.

Precision atomic-beam measurements have contributed to many theoretical and practical developments. The deuteron quadrupole moment (1939) pointed to the necessity of a tensor interaction in nuclear forces. The anomalous electron moment (1949) and the Lamb shift (1950) in the atomic-hydrogen fine structure were resolved by quantum electrodynamics. Nuclear spins (I) as well as magnetic-dipole (μ) and electric-quadrupole (Q) moments have been important in providing test information for the shell model (1949) and collective model (1953) of the nucleus. Atomic hyperfine-structure constants (dipole, a ; quadrupole, b ; and octupole, c), which describe the interaction between the electrons and the nucleus, as well as the numerous constants required to describe a molecule and its internal interactions, have contributed to the theory of atomic and molecular structure. The cesium "clock" or frequency standard (1952) represents a widespread practical application of beam technology

by using the hyperfine-structure transition at 9192.631770 Mc/sec (Ephemeris time) to regulate a quartz-crystal oscillator. Other frequency standards such as the thallium clock, ammonia maser (1954), and hydrogen maser (1960) also employ beam techniques for quantum-state selection. Very recent, high-energy nuclear accelerators have been equipped with atomic-beam sources to produce polarized protons.

Among nonresonant experiments, beams impinging on solid surfaces produce information on the wave nature of particles, work functions, and accommodation coefficients. Charge-exchange cross sections and interaction potentials are obtained from experiments with crossed beams, one neutral and one charged. Chemical reaction kinetics in isolated systems are studied in crossed beams of two reactants.

Four individuals have received Nobel Prizes for beam research: Otto Stern (1943) "for his contribution to the development of the molecular-ray method and for his discovery of the magnetic moment of the proton"; I. I. Rabi (1944) "for his application of the resonance method to the measurement of the magnetic properties of atomic nuclei"; P. Kusch (1955) "for his precision determination of the magnetic moment of the electron"; and W. E. Lamb (1955) "for his discoveries concerning the fine structure of the hydrogen spectrum."

A discussion of the energy levels of a simple atom and of one particular apparatus will illustrate the magnetic-resonance technique. An isolated atom in an external field, H , has energy states which are calculable from the Hamiltonian \mathcal{H}/h (Mc/sec) = $a \mathbf{I} \cdot \mathbf{J} + b$ (quadrupole operator $-g_J(\mu_0/h)\mathbf{J} \cdot \mathbf{H} - g_I(\mu_0/h)\mathbf{I} \cdot \mathbf{H}$, where some symbols have been defined pre-

viously, $g_J = \mu_J/(J\mu_0)$, $g_I = \mu_I/(I\mu_0)$, and μ_0 is the magnitude of the Bohr magneton. The first two terms represent the dipole and quadrupole hyperfine-structure interactions between the electrons and nucleus; the last two terms express the interaction of the electron and nuclear magnetic moments with the external magnetic field. For the simple case of $I = 1/2$, $J = 1/2$, $b = 0$, as in the ground electronic state of atomic hydrogen, the energy levels that arise from different "relative orientations" of the nuclear (I) and electronic (J) spins are shown in Fig. 1 as a function of the magnetic-field parameter $X = [(g_J + g_I)\mu_0 H]/\Delta W$. Here ΔW is the hyperfine-structure separation at $H = 0$. The levels are labeled by either the low-field quantum numbers (F, m_F) or by the high-field numbers (m_J, m_I), where $F = |I \pm J|$, and the m 's are the projections of F, I , or J along the field direction.

Of the many magnetic or electric resonance apparatuses, one specialized type has proved valuable for measuring atomic properties of both stable and radioactive isotopes. In Fig. 2, the "oven" or source, O, may take one of many forms—a microwave discharge to dissociate gaseous diatomic molecules, a closed tantalum crucible (with an exit slit) heated by electron bombardment, or one of many other devices for evaporating atoms. The atoms pass between the poles of three separate electromagnets (denoted A, C, and B, successively, from oven to detector). The inhomogeneous A and B magnets have eccentric cylindrical pole tips which produce a field gradient, $\partial H/\partial Z$. In this field, an atom experiences a force $\mathbf{F} = \mu_{\text{eff}}(\partial H/\partial Z)$, where $\mu_{\text{eff}} (- = \partial W/\partial H)$ is the negative slope of an energy level in Fig. 1. Within the homogeneous C field, a superimposed rf field induces state changes. Thus an atom which remains in a

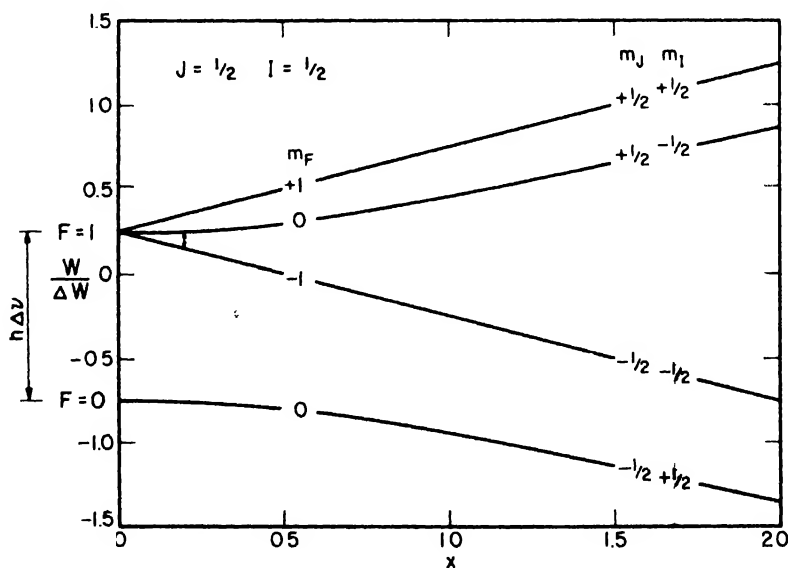


FIG. 1. Hyperfine-structure energies (ordinate) of an atom with $J = 1/2$ and $I = 1/2$ in an external magnetic field (abscissa).

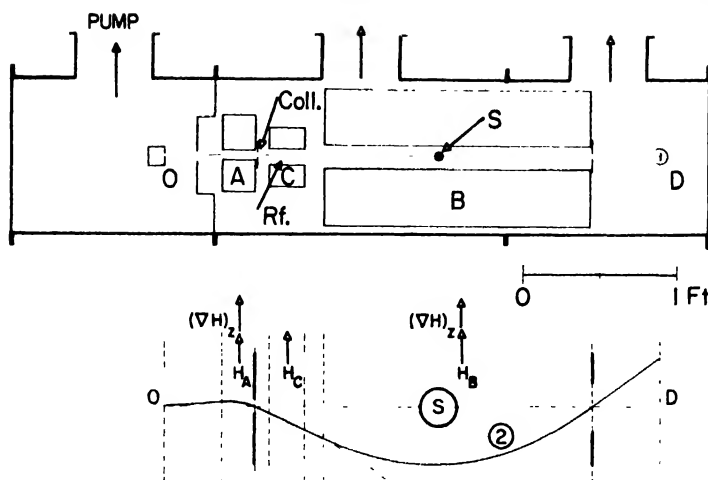


FIG. 2. Schematic of an atomic-beam, magnetic resonance apparatus.

single state [(1, 0) for example] is deflected similarly by the strong A and B magnets and follows trajectory 1 in Fig. 2. The stopwire, S, shields the detector from fast atoms and atoms with small deflections. If, in the C-field region, a transition occurs that causes the high-field slope ($\sim \mu_{eff}$) to change sign [e.g., (1, 0) \rightarrow (1, -1)], the A and B deflections are opposite, and the atom follows trajectory 2 to the detector D. A resonance is observed as an increase in beam intensity at the detector. Values of the constants in the Hamiltonian are deduced from the observed resonant frequencies of the atoms in known magnetic fields. Some detection methods in frequent use are:

(a) Deposition on a surface with subsequent assay by radioactive counting, neutron activation, or optical means (earliest detector).

(b) Ionization of alkali atoms on a hot tungsten wire, and measurement of the resulting ion current.

(c) Electron-bombardment ionization with subsequent mass analysis to discriminate against background gas ions. The beam ions are frequently counted by using electron multiplier tubes.

(d) Other detectors employing the principles of radiometers, pressure manometers, thermopiles bolometers, and changes in space charge.

HOWARD A. SHUGART

References

- Estermann, I., Ed., "Recent Research in Molecular Beams," a collection of papers dedicated to Otto Stern, on the occasion of his seventieth birthday, New York, Academic Press, 1959.
Kopfermann, H., "Nuclear Moments," English translation by E. E. Schneider, New York, Academic Press, 1958.

Kusch, P., and Hughes, V. W., in Flügge, S., Ed., "Handbuch der Physik," Vol. 37/1, Berlin, Springer Verlag, 1959.

Nierenberg, W. A., in *Ann. Rev. Nucl. Sci.*, 7, 349 (1957).

Ramsey, N. F., "Molecular Beams," London Oxford University Press, 1956.

Smith, K. F., "Molecular Beams," London, Methuen and Company, 1955.

Cross-references: ATOMIC CLOCKS, CROSS SECTION AND STOPPING POWER, MAGNETIC RESONANCE, CHEMICAL KINETICS.

ATOMIC CLOCKS

Atomic clocks make use of a property that is generally found only in systems of atomic dimensions: such systems cannot contain arbitrary amounts of energy, but are restricted to an array of allowed energy values E_0, E_1, \dots, E_n . If an atomic system wants to change its energy between two allowed values, it must emit (or absorb) the energy difference, for instance by emission (or absorption) of a quantum of electromagnetic radiation. The frequency f_{ij} of this radiation is determined by the famous relation

$$|E_i - E_j| = \Delta E = hf_{ij}$$

(h is Planck's constant). The rate of an atomic clock is controlled by the frequency f_{ij} associated with the transition from the state of energy E_i to the state of energy E_j of a specified atomic system (such as a cesium atom or an ammonia molecule). A high-frequency electromagnetic signal is stabilized to the atomic frequency f_{ij} and a frequency converter relates the frequency f_{ij} to a set of lower frequencies which then may be used to run a conventional electric clock.

The atomic frequency f_{ij} is, according to present knowledge, free of inherent errors; it is in particular not subject to "aging" since any

transition which the system makes puts it in a state of completely different energy, where it cannot falsify the measurement. Herein lies the principal advantage over other methods of time measurement. Two atomic clocks have exactly the same calibration as long as they are calibrated against the same atomic transition. Atomic readings made in Boulder, Colorado and Neuchâtel, Switzerland between 1960 and 1963 differed on the average by less than 3 msec, whereas the deviation of the astronomically measured time TU_2 from atomic time is of the order of 50 msec¹. For this reason, the atomic second was adopted as the new time unit, by the Twelfth General Conference on Weights and Measures, in October, 1964, and temporarily defined as the time interval spanned by 9 192 631 770 cycles of the transition frequency between two hyperfine levels of the atom of cesium 133 undisturbed by external fields.

The accuracy with which f_{ij} can be measured depends mainly on the degree to which the atomic resonator can be isolated from outside influences such as chemical and electromagnetic forces, and the extent to which Doppler shift can be avoided. Atomic transitions are chosen which have a minimum dependence on electromagnetic fields, and such fields as the magnetic field of the earth are carefully shielded out. Residual fields are nevertheless at present a major source of error in the most accurate clock. In many of the currently available clocks, the atomic systems are moving through vacuum with velocities of the order of 10^4 cm/sec. This motion introduces longitudinal Doppler shifts of the order $\sim 10^{-8} f_{ij}$. The transversal velocity causes only second-order shifts of the order $10^{-12} f_{ij}$. The transitions have to be observed mostly in the transverse direction, and residual longitudinal components must occur with either sign with equal probability, so as to produce, instead of shift, only a broadening that is symmetrical about f_{ij} . Whenever the systems collide with each other or are in contact with surrounding material, such as a buffer gas, they are perturbed by chemical forces and f_{ij} undergoes further shifts. Another source of error is the fact that most systems are observed only during a finite time τ . This circumstance gives rise to observation-time broadening of the order $1/\tau$ (typically, 10^3 cps). The error introduced hereby is not given by the broadening δf , but only by $\delta f/r$, where r is the signal-to-noise ratio with which f_{ij} is observed. However, if the phase shift of the observing system varies appreciably over the line width δf , additional errors known as "pulling" are introduced.

Two main problems have to be solved in atomic clocks: the atomic system has to be suspended without perturbation from chemical forces, and the frequency f_{ij} has to be measured. The most radical solution of the first problem is found in molecular-beam devices. The atomic systems are observed in free flight, and only during the relatively short transit time, so that observation-time broadening and Doppler shifts are large. Such devices have nevertheless provided the clocks with the highest accuracy so far, of the

order of a few parts in 10^{11} . The problems associated with Doppler shift and observation-time broadening can be alleviated by enclosing the atomic system in a box that is small compared to the wavelength $\lambda = c/f_{ij}$. The problem is to find atomic systems and wall materials such that the radiative process is not interrupted, and the frequency f_{ij} not appreciably influenced, during collision of the atomic systems with the walls. Instead of walls, a buffer gas can also serve as a "container" for the atomic systems. Very long observation times, giving rise to resonances of very high Q , have been produced by this technique. The accuracy of such clocks depends on the degree to which the influence of chemical forces can be neutralized. An indication of the accuracy of which such devices may ultimately be capable can be found in the fact that the measurement of resonance radiation due to Fe^{57} nuclei built into a crystal lattice has made possible frequency determinations with a relative accuracy of some parts in 10^{14} , and effects on time due to general relativity have been measured by this method. However, very high frequencies ($\sim 10^{20}$ cps) were involved, and no clock has been built for lack of a suitable frequency converter.²

The main methods of measuring f_{ij} are the following: in thermal equilibrium, the atomic systems in the lower-energy state outnumber the ones in the higher-energy state, and radiation of frequency f_{ij} induces more absorption than emission, giving rise to a net absorption signal. Several methods are known by which the population of either energy state is enhanced, giving rise to stronger absorption or emission signals. Particularly elegant is the maser principle. The higher-energy state population is sufficiently enhanced so that the emitted radiation itself can sustain the emission process. Often it is preferable to observe the absorption or emission of radiation of frequency f_{ij} indirectly. It is possible to observe microwave absorption via the change in the interaction of the absorbing atomic systems with optical radiation. Another example of indirect absorption is given by the atomic-beam method, where the microwave absorption is monitored by the deflection of the atoms from their trajectory. Indirect observation can be more economical since each microwave transition can be counted, whereas it takes several transitions to make a count by direct observation as long as the quanta to be counted carry less energy than the quanta of thermal radiation.

Several articles describing modern atomic clocks can be found in the Proceedings of the Third International Congress on Quantum Electronics, particularly in the third chapter of the first volume³.

MARTIN PETER

References

1. Bonanomi, J., Kartaschoff, P., Newman, J., Barnes, J. A., and Atkinson, W. R., "A Comparison of the TA-1 and NBS-A Atomic Time Scales," *Proc. I.E.E.E.*, **52**, 439 (1964).

2. Wertheim, G. K., "The Mossbauer Effect," *Nucleonics* (January, 1961).
3. Grivet, P., and Bloembergen, N., "Quantum Electronics," Vol. 1, Paris, New York, Columbia University Press, 1964.

Cross-references: ATOMIC AND MOLECULAR BEAMS, DOPPLER EFFECT, LASERS, MASER, QUANTUM THEORY, RELATIVITY.

ATOMIC ENERGY

The terms "atomic energy" and "nuclear energy" are used interchangeably in the contemporary literature to mean energy that originates within the atomic nucleus. Events that release atomic energy involve basic changes in nuclear structure and result in the formation of one or more different nuclides, which may be isotopes of the original atom or altogether different elements. The release of atomic energy is thus a more fundamental process than the release of chemical energy, which merely involves a regrouping of intact atoms into different molecular forms.

To date, three basic atomic energy mechanisms have been exploited in practical applications: (1) the fission of certain heavy nuclides; (2) the fusion of certain light nuclides; and (3) the process of radioactive decay. These will be discussed in the order listed.

Fission. In fission, a heavy nuclide splits into two lighter and predominantly unstable nuclides, commonly referred to as fission products, with the accompanying emission of several neutrons and the release of approximately 200 MeV of energy. Nuclides that readily undergo fission on interaction with low-energy or "slow" neutrons (< 0.5 eV) are referred to as fissile materials. There are three primary fissile materials:

(1) Uranium 235 which is a natural constituent of the uranium element and accounts for 0.71 per cent by weight of that element as found in nature.

(2) Plutonium 239, formed by neutron irradiation of uranium 238.

(3) Uranium 233, formed by neutron irradiation of thorium 232.

Uranium 238, which does not undergo fission on interaction with slow neutrons, does so on interaction with high-energy or "fast" neutrons (> 0.1 MeV). Table I lists representative fission energy distributions for the four nuclides cited.

A useful rule of thumb is that an energy release of 200 MeV per fissioning atom corresponds to an output of approximately one megawatt-day of thermal energy per gram of fissioned matter.

Practical applications of fission are based on the principle of a self-sustaining fission chain reaction, i.e., a reaction in which a neutron emitted by atom A triggers the fission of atom B, and one from atom B triggers the fission of atom C, and so on. For this to be achieved requires the assembly of a "critical mass" of fissile material, i.e., an amount sufficient to reduce the probability of neutron losses to a threshold value. The amount required depends on a number of factors, notably the concentration of the fissile material used and the composition and geometry of the reaction system.

There are two basic application concepts. One is the essentially instantaneous fission of a mass of highly concentrated fissile material in such a way as to generate an explosion force. This, of course, is what occurs in atomic weapons. Atomic explosives are also of interest in connection with peaceful uses such as large-scale excavation projects, a field of application that is currently being studied by the U.S. Atomic Energy Commission.

The other application concept is that of the controlled and gradual fission of an atomic fuel in a nuclear reactor, which may be designed for one or more of the following principal purposes:

(1) To provide fluxes or beams of neutrons for experimental purposes. This category of use includes research and materials testing reactors.

(2) To produce materials by neutron irradiation. Examples are reactors used primarily to produce plutonium for atomic weapon stockpiles or for the production of various radioisotopes for use in science and industry.

TABLE I. ENERGY DISTRIBUTION IN FISSION

Type of Energy	Quantity of Energy (MeV)			
	Slow Fission		Fast Fission	
	U ²³⁵	Pu ²³⁹	U ²³³	U ²³⁸
Kinetic energy of fission products	165	172	163	163
Kinetic energy of neutrons emitted	5	6	5	5
Instantaneous emission of gamma rays	8	7	7	7
Beta emission during fission product decay	9	9	9	9
Gamma emission during fission product decay	7	7	7	7
Total*	194	201	191	191

* Exclusive of nonrecoverable energy associated with neutrino emission during fission product decay. All numbers are rounded to the nearest integer. It should be mentioned that 8 or 9 MeV of additional energy become available in a nuclear reactor as the result of neutron capture and subsequent gamma-decay phenomena.

(3) To supply energy in the form of heat for such applications as the generation of electric power, the propulsion of ships or space vehicles, or the production of process steam.

The first demonstration of a fission chain reaction was achieved by E. Fermi and co-workers on December 2, 1942 when the world's first nuclear reactor (Chicago Pile No. 1) was successfully operated in a converted squash court beneath Stagg Field at the University of Chicago.

The most important application of fission promises to be in the electric power field. The basis for this expectation is that, if exploited efficiently, known and inferred deposits of atomic fuels represent a potential energy reserve many times larger than that of the fossil fuels (coal, oil and natural gas) on which the world's electric energy economy largely depends at present.

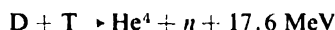
At this writing (1965), some 5000 electrical megawatts of atomic power capacity are in operation, under construction, or planned for construction in the United States. The amount of atomic electric capacity is expected to multiply more than tenfold by 1980, which would mean that atomic power would then account for 15 or 20 per cent of total U.S. electric power production. The U.S. Atomic Energy Commission has forecast that the atomic contribution may well increase to 50 per cent by the end of this century.

Fusion. Fusion is a general term for reactions in which the nuclei of light elements combine to form heavier and more tightly bound nuclei with the simultaneous release of large amounts of energy. In order for this to occur the interacting nuclei must be brought sufficiently close together to permit short-range nuclear forces to become operative. This means that one or both nuclei must be accelerated ("heated") to velocities sufficient to overcome the strong electrostatic repulsion that exists between particles having the same electrical charge. The velocities required correspond to particle "temperatures" of the order of tens or hundreds of millions of degrees, which in turn correspond to particle energies of thousands or tens of thousands of electron volts. The term "thermonuclear" reactions is reserved for fusion reactions in which both nuclei are traveling at high velocity (as distinct from reactions between an accelerated projectile particle and a static target nucleus, as in particle accelerator experiments).

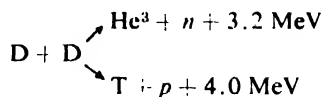
The only practical application of thermonuclear reactions developed to date is in thermonuclear weapons (so-called "hydrogen bombs") in which the energy released by a charge of fissile material serves to create the conditions required to bring about the reaction of "fusionable" materials. The first test of a thermonuclear weapon, which was the first demonstration of a man-made thermonuclear reaction, took place on October 31, 1952 at a U.S. testing site in the Marshall Islands. Peaceful uses of thermonuclear explosives are being studied and have the advantage, relative to straight fission-based explosives, that problems of radioactive contamination

are greatly reduced. This reflects the fact that the nuclides formed by fusion are stable and hence, apart from neutron activation effects, the formation of radioactive substances is limited to the fission component of the explosive.

Research has been in progress for more than a decade on techniques for controlling the fusion process as a means of supplying energy for electric power generation. The thermonuclear reactions of primary interest in this context are the deuterium-tritium reaction:



and the deuterium-deuterium reactions:



Deuterium is a stable isotope of hydrogen with a natural abundance of 0.0015 per cent. Tritium is an unstable hydrogen isotope with a radioactive half-life of 12.3 years and is produced from lithium 6 by the neutron-alpha reaction. The latter thus represents a relatively expensive "fuel" for thermonuclear reactions; however, the ignition temperature of the deuterium-tritium reaction is roughly an order of magnitude lower than that of the deuterium-deuterium reactions and the energy release is greater.

In a controlled fusion system as presently conceived, the fuel is in the form of an ionized gas, or "plasma," confined by magnetic pressure within a high-vacuum apparatus. In effect, the plasma is held in a "magnetic bottle," thereby preventing fuel particles from dissipating heat in collisions with the physical walls of the apparatus. The objective is to achieve a situation in which an adequately hot plasma of adequate density can be magnetically confined for a long enough interval of time for the desired reaction to take place. One approach is to constrict and confine a high-current discharge of fuel ions and hold the resulting dense plasma in confinement while its temperature is raised by adiabatic compression or other methods. Another approach is to accelerate fuel ions to high energies and then trap them in a magnetic field, maintaining confinement long enough for a dense plasma to accumulate.

In experiments in various experimental devices, the time-temperature-density multiple has steadily been increased; however, there is as yet no conclusive evidence that true thermonuclear conditions have yet been achieved in any laboratory. Beyond laboratory demonstration of controlled fusion per se lies the problem of demonstrating that devices can be designed to produce more power than they consume and beyond that lies the problem of demonstrating the economic feasibility of practical thermonuclear power plants. Thus, at present (1965), controlled fusion is still at the stage of basic research.

The chief incentive for thermonuclear power

development is the promise of a virtually inexhaustible energy source, assuming the ultimate use of deuterium as the primary fuel.

Radioactive Decay. As radioactive atoms undergo decay by alpha, beta or gamma emission, heat is generated by the interaction of the radiation with surrounding matter. Devices that utilize this heat to produce electricity are known as isotopic power generators. A family of such devices is being developed by the U.S. Atomic Energy Commission for specialized applications requiring from fractions of a watt to tens of watts of electricity. Thermoelectric or thermionic techniques are used to convert the heat to electricity. Costs are presently about \$10,000 per watt of electrical capacity. With further development and with volume production of isotopic "fuels," costs are expected to be reduced to the order of \$100 per watt. At present, isotopic power generators are being used on an experimental basis in a number of applications such as navigational satellites, automatic weather stations and coastal light buoys, all of which require a compact power source that can operate unattended for sustained periods (months or years). In the case of space applications, alpha-emitting radionuclides such as plutonium 238 or curium 244 are mainly used as the fuel. In terrestrial applications, the principal fuel used to date is strontium 90, a beta emitter.

JOHN F. HOGERTON

References

- Hogerton, John F., "The Atomic Energy Deskbook," New York, Reinhold Publishing Corp., 1963.
 Glasstone, Samuel, "Sourcebook on Atomic Energy," Princeton, N.J., D. Van Nostrand Co., 1958.

Cross-references: FISSION, FUSION, NUCLEAR REACTIONS, NUCLEAR REACTORS, NUCLEONICS, RADIOACTIVITY.

ATOMIC PHYSICS

It is generally accepted by those who inquire into the properties of matter that these properties can be understood in terms of small constituents called atoms. The nature and structure of atoms has been a principal subject of study by physicists and chemists during the first half of the twentieth century. As a consequence of the success of such studies, research into the structure of matter has become divided into two moderately well-separated fields. One of these concerns the further details of the structure of the atomic nucleus which is described under the terms "nuclear physics," and "high-energy physics." The other is a study of the behavior of large groups of atoms, and is generally described under the heading "solid-state physics." This article is devoted to the structure of atoms themselves, a field which, in a sense, lies between the fields of nuclear physics and solid-state physics.

The idea of atoms is very old and was known to the Greeks during the classical period. It appeared from time to time in medieval works, although the concepts expressed now seem to us

very vague. They seemed to have been based on the idea that there could be a limit to the divisibility of matter and, consequently, the idea of a final indivisible particle out of which large pieces of matter could be built.

At the beginning of the nineteenth century, it became clear that chemical reactions could be most simply explained if each chemical element is thought of as composed of very small identical atoms characteristic of the element. There thus arose a rather well-defined idea of a chemical element composed of identical atoms, as distinguished from a compound composed of groups of different atoms combined into molecules. During the latter part of the nineteenth century, the kinetic theory of gases made use of the idea of atoms, and molecules, in explaining the behavior of gases. By the end of the century only a few physicists and chemists still doubted the actual material existence and "reality" of atoms.

It is perhaps rather curious that the idea of atoms became really well established only after it became clear that the atoms were not in any true sense indivisible, but that instead they had a complex structure which could be extensively investigated. Since this investigation required equipment and methods which had been developed by physicists rather than chemists, physicists took the lead and the work has taken the name of atomic physics or the physics of atomic structure. It was inaugurated in the 1890's when J. J. Thomson first isolated and established the existence of electrons. He showed these to have only about 1/2000 the mass of the lightest known atom, hydrogen. He showed also that these electrons, as indicated by their name, carried negative electrical charges. It was later proved by R. A. Millikan that all the electronic charges are the same. Thus the identification of electrons, as small electrically charged pieces of matter, and as constituents of all matter, became firmly established.

Since it was clear that normal matter is electrically neutral, it had to be assumed that each atom contained a positive electrical charge, as well as negative electrons. J. J. Thomson developed the picture of a somewhat spherical jelly-like mass of positive electricity, in which electrons were located at various positions, and bound to them by quasi-elastic forces.

A principal means of investigating the structure of atoms has been the examination of the light emitted by the material in the gaseous state. This LIGHT is found to consist of a number of discrete wavelengths, or colors. Each of these wavelengths was associated, in the early days of this century, with a mode of vibration of the electrons in the positive jelly. In particular, H. A. Lorentz of the University of Leiden was able to show that such electrons, when placed in a magnetic field, would have their modes of vibration changed in a way that explained findings of P. Zeeman, who had made early observations of the wavelengths of the light emitted by a radiating gas in a magnetic field.

Nevertheless, the picture was unsatisfactory.

The number of different frequencies emitted by each atom was very much larger than the number of electrons which could be attributed to the atom. Furthermore, these frequencies did not show the harmonic relationship to be expected on the simple picture.

During 1910 and 1911, Sir Ernest Rutherford suggested an experiment, carried out by Geiger and Marsden, in which alpha particles from a radioactive source were scattered from thin foils. The angles at which the alpha particles were scattered were found to be such as could be best described by the close approach of a heavy positively charged particle, the alpha particle, to another heavier and more highly positively charged particle, representing the scattering atom. From the results of these experiments, Rutherford concluded that the mass in the positive charge of an atom, instead of being distributed throughout the volume of a sphere of the order of 10^{-8} cm in radius, was concentrated in a very small volume of the order of 10^{-12} cm in radius. He thus developed the idea of a nuclear atom. The atom was pictured as a small solar system with the very heavy and highly charged nucleus occupying the position of a sun, and with electrons moving around it, as planets in their respective orbits.

Although this picture of nuclear atoms served to describe the alpha-particle scattering experiments, it still left many questions unsolved. One of these questions referred to the apparent stability of the atoms. An ELECTRON moving around the nucleus would tend to emit radiation, to lose its energy, and thereby to spiral into the nucleus. Why did it not do so? Why did the atoms all seem to be quite stable, and all to be of approximately the same size, even though some contain 90 or more electrons, while hydrogen contains only one?

The first approach to a treatment of these problems was made by Niels Bohr in 1913 when he formulated and applied rules for "quantization" of electron motion around the nucleus. He postulated states of motion of the electron, satisfying these quantum rules, as peculiarly stable. In fact, one of them would be really permanently stable and would represent the ground state of the atom. The others would be only approximately stable. Occasionally an atom would leave one such state for another and, in the process, would radiate light of a frequency proportional to the difference in energy between the two states. By this means he was able to account for the spectrum of atomic hydrogen in a spectacular way. Bohr's paper in 1913 may well be said to have set the course of atomic physics on its latest path.

Out of the experimental work on the scattering of alpha particles and the theoretical work of Niels Bohr, there grew a fairly definite picture of an atom which could be correlated with its chemical properties. The chemical properties were determined in the first place by the nuclear charge. The nucleus contained most of the atomic mass and carried an electric charge equal to an

integral number of positive charges, each of the same magnitude as an electronic charge. This positive nucleus then accumulated around itself a number of electrons just sufficient to neutralize its positive charge and form a neutral atom.

The number of positive charges, or the number of negative electrons around the nucleus, was designated as the atomic number of the atom. These showed a close parallelism with atomic weights, and an even closer parallelism with the arrangement of atoms in the periodic system. Through the formulation of a number of rules based on Bohr's picture of quantized orbits, the periodic system of the elements could be understood. Hydrogen was given one electron, and helium two. The two electrons in helium constituted a "closed shell" which exhibited almost perfect spherical symmetry and chemical inactivity. Since the first two electrons constituted a closed shell the third electron in the case of lithium gave the atom properties very similar to those of hydrogen. Outside of the first closed shell could be placed eight more electrons to form a second closed shell in neon. The four outer electrons necessary for the atom of carbon provided a description of the chemical properties of that ubiquitous element, and the one electron short of a closed shell in the case of fluorine gave some understanding of fluorine's vigorous chemical activity.

Thus, during the years after 1913, the feeling grew that the chemical properties of atoms could be pretty well understood. The idea that there were undiscovered elements, as indicated by gaps in the periodic system, was reinforced. Systematic search has now discovered them.

However, it was not until 1925 that Bohr's ideas were developed into a mathematical form, complete enough and precise enough to permit their general application, under the name of quantum mechanics. This development, associated with the names of Dirac, Heisenberg, Schrödinger, provided the basic laws which permit, in principle, the complete and quantitative description of an atom consisting of a heavy positively charged nucleus, and surrounded by enough electrons to make the whole system electrically neutral.

The principal feature of quantum mechanics is that it ascribes to electrons something of the nature of waves. It requires that calculations of the motion of an electron in the neighborhood of a nucleus be made in terms of wave motion rather than the motion of particles. One has immediately the explanation of the stability of atoms. The waves representing the electrons cannot collapse to points, or to the small size of the nucleus, without involving extremely large amounts of "kinetic" energy. The ground state of an atom is determined as the balance between the attraction of the nucleus for the electron wave and what might be called the elastic resistance to compression of the wave itself.

Other states of the atom may be pictured as states of vibration of the electronic wave around the nucleus. The state with a minimum number

of nodes is the ground state, and states with more complicated vibration forms represent higher energies and are called excited states. The excited states are in general unstable in the presence of an electromagnetic field. Most excited states decay into states of lower energy with a lifetime of about 10^{-8} second. In the process, the atom radiates an electromagnetic wave of a rather sharply defined frequency, equal to the energy difference between the two states divided by Planck's constant h .

The details of these states and of their rates of decay can all be calculated in principle, although the calculation involves as much mathematical difficulty as any calculation of the behavior of a large number of particles interacting with each other.

One of the properties of electrons that became evident during the study of the optical spectra of atoms was that of ELECTRON SPIN. The suggestion was made by Uhlenbeck and Goudsmit in 1925 that one of the features of such spectra could be understood if each electron had associated with it a quantity called spin, which is similar in many ways to angular momentum. Each electron also has a certain magnetic moment which affects the energy in the presence of a magnetic field. This property also has been incorporated into the wave ideas of quantum mechanics, and may be thought of as leading to two similar waves existing at the same time but each corresponding to a different "spin."

It is widely believed that the subject of atomic physics is closed in principle except for detailed and complicated calculations which can be made with computing machines. On the one hand, physicists have gone on to investigating the detailed properties of the atomic nucleus. Most of these properties have little effect on the atoms themselves. On the other hand, the properties of atoms in combination, and the way in which they form large pieces of solid matter, has led to an extensive and comprehensive field known as solid-state physics. These two directions lead to the exciting frontiers of physics in the second half of the twentieth century.

W. V. HOUSTON

References

- Fano, U., and Fano, L., "Basic Physics of Atoms and Molecules," New York, John Wiley & Sons, 1959.
Blackwood, Osgood, and Ruark, "An Outline of Atomic Physics," New York, John Wiley & Sons, 1955.

Cross-references: ELECTRON; ELECTRON SPIN; ELEMENTS, CHEMICAL; MOLECULES AND MOLECULAR STRUCTURE; NUCLEAR STRUCTURE; PERIODIC LAW AND PERIODIC TABLE; QUANTUM THEORY; SPECTROSCOPY; ZEEMAN AND STARK EFFECTS.

ATOMIC SPECTRA

Fundamental Facts. Light from electric discharges in gases shows *line spectra* due to free atoms excited by electron collisions. Noble gases

and metal vapors produce almost pure atomic spectra, while discharges in molecular gases show both molecular *band spectra* and atomic line spectra. Some spectral lines can also be observed in absorption when white light is made to pass through the gas into a spectroscope. Under high spectroscopic resolution, all lines are found to have nonzero width. This is due to random motion (DOPPLER EFFECT) and disturbing influences of neighboring atoms, molecules, ions or electrons (pressure broadening); but even after allowance for these effects, a spectral line has a definite, generally very small, width due to radiation damping (natural width). Precision measurement of wavelengths or resolution of very fine structures requires light sources giving narrow lines—discharges at low gas and current density and low temperature or even atomic beams at right angles to the line of sight. *Continuous* atomic spectra are generally weak under laboratory conditions; in emission, they are due to recombination of an electron with a positive ion; in absorption, to the reverse process of photoionization.

The term atomic spectra includes positive ions, with the following terminology: spectrum of Na, *arc spectrum*, NaI; of Na⁺, Na⁺⁺, ...; first, second, ... *spark spectrum*, or NaII, NaIII, ... Spectra of highly ionized or *stripped* atoms are important in astrophysics and occur in high-temperature plasmas. Systems with the same number of electrons, such as Na, Mg⁺, Al⁺⁺ show marked similarities and are called *iso-electronic* sequences.

In the ultraviolet, visible or infrared, the spectroscope, in the form of a grating or interferometer, measures primarily the wavelength (λ) of the spectral lines. It is generally expressed in angstrom units (\AA) defined as 10^{-8} cm or, by recent international convention, the fraction $1/6056.12525$ of the wavelength of a line of the isotope 86 of krypton, in air under standard conditions. The wave number ($\tilde{\nu}$ or σ), the reciprocal of the wavelength in vacuo, is measured in cm^{-1} or kayser (K), or in millikayser (mK). The frequency ν is derived by multiplying by c , the velocity of light in vacuo; in the range of microwaves and radio frequencies, ν is measured directly.

In contrast to frequencies, intensities of lines are strongly dependent on experimental conditions, and special experiments are required for deriving quantities expressing the strength of a line as a characteristic constant of the atom. This can be defined in various forms; the *f*-value is a number giving the ratio of the absorptive or dispersive power of the line to that of the classical, harmonic electron oscillator of the same frequency; the transition probability or Einstein *A*-value is the probability, per second, of an excited atom emitting a light quantum.

Hydrogen-like Spectra. The spectra of atoms containing one electron only (H, He⁺, Be⁺⁺, ...) are very simple if the fine structure is disregarded; they form the basis of the classification and theory of atomic spectra. Balmer's empirical

discovery of a numerical relationship between the wavelengths of the visible hydrogen lines led to a formula expressing the wave numbers of all hydrogen-like spectra by one constant R , the charge number Z ($=1$ for H; $=2$ for He^+ , ...) and two integral numbers n, n' ($=1, 2, 3, \dots$):

$$\tilde{\nu} = Z^2 R(1/n^2 - 1/n'^2) = T_n - T_{n'} \quad (1)$$

A series arising from a sequence of values n' is characterized by regularly decreasing spacings and intensities of the lines towards increasing wave numbers. Substitution of $n=1, 2$ and 3 in Eq. (1) with $Z=1$ gives the Lyman, Balmer and Paschen series, in the ultraviolet, visible and near infrared respectively. The wavelengths of the Balmer lines $\text{H}_\alpha, \text{H}_\beta$ and H_γ ($n'=3, 4, 5$) are 6562.8, 4861.3 and 4340.5 Å. The Lyman α line, the resonance line of hydrogen, has the wavelength 1215.7 Å.

The relation of Eq. (1) can be derived theoretically by applying nonrelativistic quantum theory to a model consisting of a point electron of mass m and charge $-e$ and a fixed point nucleus of charge Ze . In the *Bohr-Sommerfeld* theory, this is done by imposing quantum conditions on the classical orbit of the electron; in the more rigorous Schrödinger theory, by solving the wave equation with the assumption of constant energy E . Provided $E < 0$ (bound state), it assumes discrete values, those of the stationary states of motion or the *eigenvalues* of the wave equation. Emission and absorption arise from transitions between two energy levels $E_n, E_{n'}$, with the frequency of the light given by

$$\nu_{n,n'} = (E_{n'} - E_n)/h \quad (2)$$

where h is Planck's constant. Equation (1) is a special case of Eq. (2), with $E_n = -h^2 Z^2 R/n^2$. Allowance for the motion of the nucleus of finite mass M causes R to differ slightly for different M ; it is given by $R_\infty/(1 + m/M)$ where $R_\infty = 109737.3 \text{ cm}^{-1}$.

The solution of the SCHRODINGER EQUATION for a mass point in space leads to 3 quantum numbers. In polar coordinates, with a force derived from a central potential $V(r)$, the quantum numbers n_r, μ and m give the numbers of nodes of the wave function in the range of the coordinates r, ϑ and φ . If a new set of quantum numbers is defined by $l = |m| + \mu, n = n_r + l + 1$, these are found to determine the values of the following constants of the motion: the z -component of the angular momentum is $L_z = m\hbar$, where $\hbar = h/2\pi$, the square of the absolute value is $|L|^2 = l(l+1)\hbar^2$ and the energy E depends on n and l only. For the special case of the Coulomb field $V \sim 1/r$, E depends on n alone: $E_n = \text{constant}/n^2$. An energy level E_n has to be considered as consisting of a number g of states of different l and m . This situation is described as *degeneracy*, and g is the statistical weight of the level. The degeneracy in m is due to the central symmetry of the force field and occurs in all atoms in the absence of external fields. The degeneracy in l is peculiar to the Coulomb field in nonrelativistic treatment.

Alkali-like Spectra. The spectra of the alkali atoms and their isoelectronic ions ($\text{Li}, \text{Na}, \dots, \text{Be}^+, \text{Mg}^+, \dots$) show lines arranged in series similar to those of hydrogen. Their wave numbers can be represented by empirical relations which are generalizations of Eq. (1). Series of term values T_n can be defined in such a way that $T_n \rightarrow 0$ for $n \rightarrow \infty$, and the observed wave numbers are equal to term differences $T_n - T_{n'}$ (Ritz combination principle). In contrast to Eq. (1), however, there are several series of terms, so that apart from n , a second index number l has to be introduced. The term values $T_{n,l}$ can then be identified with quantized values of $-E/ch$ and the index numbers n and l with quantum numbers, as implied by the letters chosen, if we assume that in these atoms one electron moves in a central force field different from a Coulomb field. This *valency* or *optical* electron has to be imagined as more loosely bound than the others and moving in the field of electrostatic attraction by the core consisting of the nucleus of charge Ze and the remaining $Z-1$ electrons. At large distances from the core, the field is like that caused by a single charge e as in hydrogen. At smaller distances, the optical electron penetrates into the electron cloud of the core and experiences an increased attraction. This picture leads to a qualitative understanding of the term diagrams and spectra of alkali atoms as exemplified for Na in Fig. 1, where terms with $l=0, 1, 2, 3, 4, \dots$ are conventionally described as *S, P, D, F, G, \dots* terms. For any given n , an energy level is the further below that of hydrogen (the term value the larger) the smaller l

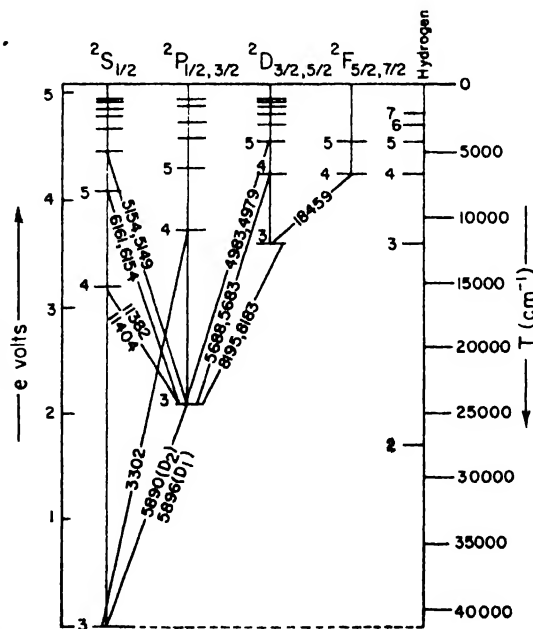


FIG. 1. Term diagram of Na. Approximate wavelengths in angstroms. The doublet splitting of the terms is not shown.

becomes, because a smaller angular momentum decreases the centrifugal force and brings the electron closer to the core. Emission or absorption of radiation according to Eq. (2) does not occur for all pairs of levels but is subject to the *selection rule* $\Delta l = \pm 1$. Transitions from P levels to the lowest S level form the *principal series*. Its lines can also be observed in absorption since the lower level is the ground level; the first member is the well-known yellow *resonance* line. Transitions from D or higher S levels to the lowest P level form the *diffuse* and *sharp* series, those from F levels to the lowest D level, the *Bergmann* or *fundamental* series.

A feature that cannot be explained by this model is the *doublet* structure, a doubling of all except the S levels. It is due to the fact that the electron possesses a *spin*, i.e., an intrinsic angular momentum S of fixed absolute value given by $|S|^2 = s(s+1)\hbar^2$, where $s = \frac{1}{2}$, and connected with a magnetic dipole moment μ . The interaction of μ with the magnetic field due to the orbital motion, the *spin-orbit coupling*, causes any one energy level of given n and l (except for $l = 0$) to split into two levels. They are characterized by a new quantum number j associated with the total angular momentum resulting from vector addition of L and S . It can have the two values $j = l \pm s = l \pm \frac{1}{2}$. Optical transitions are subject to the selection rule $\Delta j = \pm 1$ or 0. The width of the doublet splitting increases with core penetration and thus with decreasing l ; it is most prominent in P terms. It decreases rapidly with increasing n and increases from Li to Cs. In the term symbol, j is written as suffix and the doublet character is indicated by superscript 2. For example, two transitions forming the yellow resonance doublet of Na are written $3^2S_1 - 3^2P_1$ and $3^2S_1 - 3^2P_2$. The absolute value of n ($= 3$ in this case) can be deduced by comparison with hydrogen (Fig. 1).

Since for all term values $T_n > 0$ for $n \rightarrow \infty$, the extrapolation of a series $T_n = T_{n'}$ for $n' \rightarrow \infty$ gives the term value T_n . If this is the ground term, chT_n is the ionization energy of the atom. A convenient conversion formula is $8066 \text{ cm}^{-1} = 1 \text{ eV}$.

Under high resolution, hydrogen-like spectra are found to have a rather complex structure known as *fine structure* (the same name is often applied to the much wider doublet or multiplet structures); it can be explained by a relativistic velocity dependence of the electron mass removing the degeneracy of states of different l , and by magnetic spin-orbit coupling causing a doublet splitting. In fact, these two effects are related since the spin itself is relativistic in origin. The theory gives the result that a hydrogen level of given n depends on j only, so that, e.g., $n = 2$, $l = 0$, $j = \frac{1}{2}$ should coincide with $n = 2$, $l = 1$, $j = \frac{1}{2}$. In fact, such terms show a small difference known as *Lamb shift*. Its existence can be explained by quantum electrodynamics.

The treatment of the core in alkali atoms as providing a central force field is theoretically justified by the *Pauli principle*. In spite of the mutual

electrostatic interaction between electrons, it is permissible to ascribe quantum numbers n and l to them individually. This is implied in the nomenclature describing, for example, a configuration of two electrons with $n = 2$, $l = 1$ by the symbol $2p^2$. The Pauli principle allows not more than $2(2l+1)$ electrons to have the same quantum numbers n and l (*equivalent* electrons). When this number is reached, a *closed sub-shell* is said to be formed, characterized by spherical symmetry of charge distribution and vanishing total angular momentum. Thus the two $1s$ electrons in He, the further two $2s$ electrons in Be and the further six $2p$ electrons in Ne form closed subshells. In He and Ne, all electrons permissible for a given n (1 or 2, respectively) have been filled in and the configuration is described as a *closed shell*.

Other Simple Spectra. Helium, and also Li^+ , Be^{++} , \dots , have two electrons, and the atoms in the second column of the periodic table Be, Mg (also B^+ , Al^+ , \dots) have two electrons outside closed shells. The analysis of the spectra leads to two systems of terms, *singlets* and *triplets*, with only weak intercombinations; the S terms are single also in the triplet system. These facts can be formally described, in analogy to alkali spectra, by vector addition of the two spins to form the two possible resultant spin quantum numbers $S = 0$ and 1. The first alternative produces singlets ($j = l$), the second triplets ($j = l \pm 1$, $l, l+1$) unless $l = 0$. This implies a strong interaction forming the resultant S of the two spins and a weaker interaction forming j . The latter interaction is the same magnetic spin-orbit interaction that causes doublet splitting in alkali atoms, but the former is, in a less obvious way, due to the electrostatic repulsion between the two electrons. In all the terms concerned, the symbol for the configuration shows only one electron to be excited, e.g., $1s2p$ in He, but owing to the identity of the electrons, it is not possible to attribute the excited state $2p$ to one particular electron. The situation is analogous to that of two identical, coupled, linear oscillators showing two normal modes of vibration, each involving both oscillators in a *symmetrical* and *antisymmetrical* way. Application of Pauli's principle to the wave mechanical description of this two-electron system leads to two energy states, one with parallel and one with antiparallel spins, $S = 1$ and 0 respectively. Elements in the subgroup (Zn, Cd, Hg) show similar singlet and triplet spectra; in the heavier elements, however, the magnetic spin-orbit interaction is no longer weak compared with the electrostatic repulsion, and the division into singlet and triplet terms has a very restricted meaning.

In the elements of the third column, B, Al, Ga, In, Tl, the single electron outside a closed subshell produces doublet spectra, but in contrast to the alkali spectra, the ground term is a P term.

Complex Spectra. If more than one electron outside closed subshells have values of $l > 0$, the spectrum and the term structures are more

complex. Very often, especially in low-lying levels, a description in terms of the Russell Saunders coupling scheme (L, S coupling) is possible because the electrostatic interaction predominates over the magnetic spin-orbit interaction. As a result, we can define terms, each of which is characterized by a set of values L, S (or a multiplicity $2S + 1$), and each term is split into levels, each characterized by a value of j , the highest of which is equal to $L + S$. The classification and terminology are obvious generalizations of those for two-electron spectra. For configurations of 3, 4, 5 electrons the possible multiplicities are respectively: doublets and quartets; singlets, triplets and quintets; doublets, quartets and sextets, etc. The strongest lines arise from transitions between terms of the same multiplicity. Such line multiplets are often recognizable by their characteristic groupings of the lines and their intensity ratios. The level spacings are governed by the *interval rule*: they are in the ratio of the j values, e.g., the levels of $^4D_{7/2, 5/2, 3/2, 1/2}$ have spacings in the ratio 7:5:3.

Other, often much more complex forms of coupling occur, especially in the higher levels, and L and S then lose their meaning. One important property of any level which always remains well defined is the *parity*, and the Laporte rule states that even terms combine only with odd terms and vice versa. If the configuration is defined, a level is even or odd if Σl is even or odd. Levels can often be described as mixtures of several terms or configurations, but only of the same parity.

Hyperfine Structure (hfs) and Isotope Shift. These structures are usually of the order of fractions of 1 cm^{-1} and generally require interferometric methods for their study. Hyperfine structure is primarily due to the magnetic interaction of the nuclear magnetic moment μ_N with the field produced by spins and orbital motions of the electrons. A level of given j splits into hyperfine levels, each characterized by a quantum number F , where F can assume the values $j + I, j + I - 1, \dots, |j - I|$. The nuclear spin I is a characteristic property of each nucleus and has integral or half-integral values for even or odd values of the atomic number. The structure of hyperfine multiplets is similar to that of fine structure multiplets, with F, I, j taking the place of j, S, L . However, there is often also an electrostatic interaction between the electrons and the nucleus if the nuclear charge distribution has no spherical symmetry. Deviations from the interval rule in hyperfine multiplets have led to the discovery of such nuclear deformations described mainly by the *quadrupole moment* Q . Hyperfine structures in ground states can be measured very accurately by methods of atomic beam resonance. Hyperfine structure studies lead to values of I , approximate values of μ_N and Q , and accurate values for the ratio of the two latter for different isotopes.

For different isotopes of an element, the spectral lines—or the centers of gravity of their hyperfine multiplets—are often displaced against

each other. This *isotope shift* is due to two causes: in the light elements, it is due to the difference in nuclear mass, though this *mass effect* is not as easy to calculate as that causing the difference in the Rydberg constant in hydrogen-like spectra. In the heavier elements, differences in size and shape of isotopic nuclei can cause quite appreciable isotope shifts. Studies of this *volume* or *field effect* have contributed to our knowledge of nuclear volumes and the dependence of nuclear deformations on neutron numbers.

Magnetic and Electric Effects. The effects caused by magnetic fields play a great part in research on atomic spectra, particularly in magnetic resonance methods. An external magnetic field removes the degeneracy due to the spherical symmetry of atomic force fields and causes the energy to depend on the magnetic quantum number m . This leads to the formation of the *Lorentz triplet* or *normal Zeeman effect* in singlet spectra and to the more complex structures of *anomalous Zeeman effects* in multiplet lines. Zeeman effects in hyperfine structures are especially important for the determination of nuclear spins. The *Stark effect*, due to electric fields, is of somewhat less importance. It causes the energy to depend on $|m|$ only, thus not removing the degeneracy completely. While magnetic splittings are proportional to the field strength, Stark splittings are generally proportional to the square of the electric field.

H. G. KUHN

References

Textbooks

- Herzberg, G., "Atomic Spectra and Atomic Structure," New York, Prentice Hall, 1937.
Kuhn, H. G., "Atomic Spectra," London, Longmans and New York, Academic Press, 1962.

Mathematical Texts

- Condon, E. U., and Shortley, G. H., "Theory of Atomic Spectra," Cambridge, The University Press, 1935.
Slater, J. C., "Quantum Theory of Atomic Structure," New York, McGraw-Hill Book Co., 1960.

Tables

- Harrison, G. R., Massachusetts Institute of Technology Wavelength Tables, New York, 1939.
Moore, C. E., "Atomic Energy Levels," Vols. I, II and III, *Natl. Bur. Std. Circ.* 467 (1949–1958).

Cross-references: BAND SPECTROSCOPY, SCHRÖDINGER EQUATION, SPECTROSCOPY, ZEEMAN AND STARK EFFECTS.

AUGER EFFECT

Definition and History. The *Auger effect* is the filling of an electronic vacancy in the atom by one electron from a less tightly bound state, with the simultaneous emission not of a photon but of a second electron from another less tightly bound state.

Following experiments by Barkla (1909) and Sadler (1917) in which the number of characteristic K x-rays emitted by material absorbing higher-energy x-rays appeared to be substantially less than the number of x-rays absorbed in the K shell, Kossel (1923) suggested that the remaining vacancies might be filled by a radiationless transfer of the excess energy to an emitted electron. This interpretation was reiterated by Barkla and Dallas (1924), who observed an increase in the number of electrons emitted when x-rays were absorbed. Wilson (1923) had observed in a cloud chamber, simultaneous ejection of two electrons from the same atom. It remained for Auger (1925, 1926) to make systematic investigations of this phenomenon in argon. The effect has since been called the *Auger effect*, and the ejected electrons have been called *Auger electrons*.

Principal Features. Auger showed that

- (1) The photoelectron and its Auger electron arise at the same point.
- (2) The Auger-electron track length is independent of the wavelength of the primary x-rays, but the photo-electron track length increases with x-ray energy.
- (3) The direction of ejection of the Auger electron is independent of that of the photoelectron.
- (4) Not all photoelectron tracks show a coincident Auger track.

Filling of vacancies by the Auger effect can occur for any vacancy for which there are two electrons in the atom sufficiently less tightly bound than a net positive energy is available for the ejected Auger electron. Because photon emission is more easily detected and has played such an important role in the development of quantum theory, it is not generally realized that Auger emission is much more probable. Only for vacancies in the K shell in atoms with atomic number above 32 and in vacancies in the outer two electron states of an atom does photon emission dominate. The Auger effect also occurs after capture of a negative meson by an atom. As the meson changes energy levels in approaching the nucleus, the energy released may be either emitted as a photon or transferred directly to an electron which is emitted as a fairly high-energy Auger electron (keV for hydrogen, MeV for heavy elements). Finally we note that each Auger process increases the positive ionization of the atom by changing one initial vacancy into two final vacancies.

Energy Spectra of Auger Electrons. *Auger spectroscopy* is the measurement of the number, energy, and intensity of lines present. The spectrum of Auger electrons resulting from a given vacancy is more complex than the corresponding photon spectrum. The energy of the Auger electron resulting from the filling of a vacancy V of energy $E(V)$ by production of vacancies X_i and Y_j of energies $E(X_i)$ and $E(Y_j)$ is

$$E(V - X_i Y_j) = E(V) - E(X_i) - E(Y_j) - \Delta E_{X_i Y_j}$$

where $\Delta E_{X_i Y_j}$ can be interpreted as either the

increase in binding energy of the Y_i electron due to an X_i vacancy, or vice versa. Exact calculation of the number of possible Auger transitions, their energies, and their relative probabilities necessitates the use of a relativistic intermediate-coupling theory. In the above notation, X and Y refer to the total quantum number of a group of levels, and i and j to the individual substates within the group. For most Auger spectra a relativistic single-electron approximation (jj -coupling limit) gives the number of transitions and, with an empirical $\Delta E_{X_i Y_j}$, the energies of the lines to a high degree of accuracy, inasmuch as the lines not accounted for in this approximation are weak and have energies almost equal to the strong lines. The Auger electron described above would be called a $V - X_i Y_j$ electron. The spectrum from an initial vacancy of total quantum number n consists of three well-separated groups which can be characterized as $n - (n + 1)(n + 1)$, $n - (n + 1)(n + m)$, and $n - (n + m)(n + m)$, where $(n + m)$ represents all electrons with quantum numbers equal to or greater than $n + 2$. Thus for a K vacancy we have K-LL, K-LX, and K-XY groups, where X and Y stand for all electrons of total quantum number 3 or greater. For an L_3 vacancy we have L_3 -MM, L_3 -MX, L_3 -XY, and similarly for L_1 and L_2 . But for the L_3 shell the groups do not appear to be well separated, as they do for the K shell, because of the overlapping L_1 and L_2 spectra (for example L_1 -MM frequently overlaps L_3 -MX, and L_2 -MM usually falls partly between L_3 -MM and L_3 -MX). Each of these major groups contains many lines even in the jj -coupling limit. Thus the K-LL group contains six major lines K- $L_1 L_1$, K- $L_1 L_2$, K- $L_2 L_2$, K- $L_1 L_3$, K- $L_2 L_3$, K- $L_3 L_3$. Except for the K-LL group, knowledge of Auger spectra is very rudimentary for almost all elements.

Relative Probability of Auger and x-ray Emission. The evaluation of the relative probability of x-ray and Auger emission for different initial vacancies, and the determination of the relative intensities of various Auger lines constitute one of the most important aspects of Auger effect research.

The *fluorescence yield*, ω_i , for any initial vacancy i is defined as the fraction of vacancies filled by emission of *photons*. The *Auger yield* is defined correspondingly as the fraction filled by emission of *Auger electrons*. The Auger yield is divided into two parts, one (denoted by a_i) which transfers the vacancy to a level with a higher total quantum number, and the other (denoted by f_{ij}) which transfers the vacancy to a lower-energy vacancy with the same total quantum number. The latter process is called the *Coster-Kronig effect*.

For the K shell the following equation holds

$$1 = \omega_K + \alpha_K$$

Although in principle $\omega_K(Z)$ should be readily calculable, the number of (frequently relativistic) electron wave functions which must be known for each Z , and the number of permutations and

combinations of these functions which must be handled in order to calculate the individual probability of every line and, by summing, the total K Auger probability per unit time, present a formidable problem even with the aid of a sophisticated computer. The K fluorescence yield, ω_K , is therefore normally taken from experiment. Between $Z = 23$ and $Z = 90$, both ω_K and a_K are known to a few per cent or better. Below $Z = 23$, ω_K becomes so small that experiments sometimes disagree by a factor of 5. Above $Z = 90$, a_K is very small and is not known to better than a factor of 2. The following semi-empirical equation due to Burhop, with constants by Bergstrom, gives values correct to a few per cent between $Z = 23$ and $Z = 57$, and less accurate values outside these limits:

$$(\omega_K/1 - \omega_K)^{1/4} = 6.4 \times 10^{-2} + 3.40 \times 10^{-2} Z - 1.03 \times 10^{-6} Z^3.$$

Considerable experimental data exist in the literature for the relative intensities of the individual $V - X_i Y_j$ lines, especially for the K-LL group. Relativistic jj -coupling theoretical calculations for the latter group for atomic number 80 are in good agreement with experiment.

For the L shell in the jj -coupling limit the following equations hold:

$$\begin{aligned} 1 &= \omega_3 + a_3 \\ 1 &= \omega_2 + a_2 + f_{23} \\ 1 &= \omega_1 + a_1 + f_{12} + f_{13} \end{aligned}$$

In addition, the average L-fluorescence yield $\overline{\omega_L}$ for an atom with an L vacancy, having probability n_1 , n_2 , and n_3 of being in each of the three subshells, is

$$\overline{\omega_L} = 1 - \overline{a_L} = n_1(\omega_1 + f_{12}\omega_2 + f_{13}\omega_3 + f_{12}f_{23}\omega_3) + n_2(\omega_2 + f_{23}\omega_3) + n_3\omega_3$$

The values of the nine L-shell constants are much less well known than the two constants for the K shell. Below atomic number 50, the ω_i are all less than 10 per cent and appear to be less than 50 per cent for all elements. The f_{ij} go through sudden changes in value as certain transitions become energetically possible or impossible. For example, $L_1 - L_3M$ transitions are energetically forbidden for Z from 50 to 73.

The $\overline{\omega_L}$ and $\overline{a_L}$ clearly depend on the type of excitation. Although initial interest in the Auger effect arose through creation of vacancies by ejection of photoelectrons by x-rays, the strong recent upsurge of interest is the result of the studies of vacancies created in the decay of certain radioactive species. Internal conversion, shake-off accompanying beta-decay, and electron capture all create vacancies. In fact, the only method of studying the relative probability of orbital capture in the various shells and subshells of a nuclide involves the study of the x-ray or Auger spectrum of the product nuclide. The interpretation of these spectra necessitates the knowledge of the constants discussed above.

The importance of the Auger effect for radioactive nuclei, as well as the almost 100 per cent probability of Auger and Coster-Kronig emission for the M and higher levels, is indicated by measurement of the total charge accumulated by certain radioactive nuclei. For example, a vacancy produced in xenon $131m$ by internal conversion gives rise in some cases to as many as 21 Auger processes, leaving a xenon ion with a charge of $-22e$.

S. K. HAYNES

References

- Burhop, E. H., "The Auger Effect and Other Radiationless Processes," Cambridge, The University Press, 1952.
Listengarten, M. A., "The Auger Effect," (a Review), *Bull. Acad. Sci. USSR*, **24**, No. 9, 1050 (1960).

Cross-references: ATOMIC PHYSICS, PHOTOELECTRICITY, X-RAYS.

AURORA AND AIRGLOW

The visual aurora consists of luminous forms (arcs, rays, bands) in the night sky, usually confined to high latitudes, and based in the ionospheric E region (see IONOSPHERE). The airglow consists of a faint relatively uniform luminosity which is world wide in occurrence and, except under exceptional conditions, can only be observed instrumentally. The distinction between faint aurora and bright airglow in auroral regions is not clear.

The luminosity arises from emissions of the atmospheric constituents in the atomic, molecular or ionized forms. Table I shows the chief emissions in the visible region, with approximate intensities in Rayleighs for a bright aurora and temperate latitude airglow. There are many other emissions in the infrared and ultraviolet. In bright aurorae the colors can be seen visually; faint aurorae appear greyish white since the color vision threshold (except in the red) is above the visual threshold.

An auroral arc is a narrow horizontal band of light up to hundreds of kilometers long (usually geomagnetic east-west). The term arc derives from its appearance from the earth's surface due to perspective. A band is a portion of an arc showing distortion normal to its length. Auroral rays have been likened to searchlight beams; they lie along the geomagnetic field direction and may be several hundred kilometers long. Arcs and bands may be homogeneous or rayed.

Isolines of auroral occurrence are approximately centered on the geomagnetic poles. The auroral zones are defined as the regions of maximum occurrence. They are roughly circular with a radius of approximately 23° of latitude. The northern auroral zone reaches its lowest geographic latitude over eastern Canada; the southern, over the ocean south of Australia. At times of geomagnetic disturbance the aurora

TABLE 1. CHIEF AURORAL AND AIRGLOW EMISSIONS IN THE VISIBLE REGION*

Emission	Spectral Region or Wavelength	Approximate Height (km)	Approximate Intensity (Rayleighs)	
			Bright Aurora**	Nightglow
[OI]	5577Å	90-110	100,000	250
	6300,6364Å	160	50,000	150
NII	Blue to red		25,000	
H (Balmer Series)	Red, blue	E layer	1,000	
N ₂ (1st Positive)	Red	D layer	50,000	
(2nd Positive)	Violet	D layer	100,000	
N ₂ ⁺ (1st Negative)	Blue-violet		165,000	
O ₂ ⁺ (1st Negative)	Red-yellow	D layer	10,000	
NaI	5890,5896Å	80-90		100 (winter) 20 (summer)
OH	Red-yellow	60-100		100
O ₂ (Herzberg)	Blue-violet	90-100		15

* Adapted from Chamberlain, "Physics of the Aurora and Airglow," New York and London, Academic Press, 1961 to which reference should be made for information including ultraviolet and infrared emissions.

** International Brightness Coefficient III (brightness of moonlit cumulus clouds).

NOTE. Emissions are highly variable or absent with type and latitude of aurora. Heights are given only when well-defined.

appears at lower latitudes and in very great magnetic storms may be observed in the tropics. The frequency of occurrence of aurorae at lower latitudes correlates with the cycle of solar activity.

Two pieces of very recent work have thrown doubt on the significance of the idea of the auroral zone. The first is the discovery of a relatively uniform auroral glow over the polar cap, extending through and beyond the classical auroral zone, on which auroral forms appear as brighter patches, which are visible merely because of contrast with their surroundings. The second is the suggestion of a local time dependence in the daily maximum of auroral occurrence which is at about 68° geomagnetic at midnight and 75 to 80° at noon. An inner auroral zone at 75 to 80° geomagnetic could possibly explain the observed results; the evidence is inconclusive.

Many auroral forms are probably caused by the precipitation of particles (mainly electrons) into the ionosphere. Their origin is obscure, but recent investigations suggest that they are derived from the outer regions of the magnetosphere, and are accelerated and precipitated in an irregular manner on the high latitude side of the outer radiation belt through some mechanism (e.g., turbulence) which is probably related to the solar wind. It is doubtful if precipitation of trapped particles from the outer radiation belt causes aurora directly, except in great magnetic storms (see RADIATION BELTS).

A strong ionospheric current system is seated approximately in the classical auroral zone, but the detailed relation between aurorae and the electric currents is obscure.

The so-called radio aurora signifies the ionization in the E layer that is associated with magnetic disturbances, and gives rise to characteristic type radio reflections in the VHF (30 to 300 Mc/sec) band and less often at higher frequencies.

It has often been suggested that radio aurora may be identified with the optical aurora, but little real evidence exists for this. It is thought that the correspondence between them may be only random.

The chief characteristic of the ionization is that it is aligned along the earth's magnetic field, the size of the irregularities ranging from meters to kilometers in length. The mechanism producing it is obscure; wind shears and particle precipitation probably contribute.

The pattern of ionization usually shows a systematic movement which in and below the auroral zone is statistically very similar to the ionospheric disturbance current system, but there are difficulties in interpreting the movement as that of the electrons in the current system. Other interpretations are that the movement is that of the ionizing source, or even sound waves.

Frequently electromagnetic noise emission (hiss), centered around 8 kc/sec, is observed in association with aurora. Recent satellite observations have sometimes shown detailed correspondence between electron precipitation, auroral light intensity, and hiss, but at other times the correlation is poor. Theories of this noise emission all consider the interaction of a stream of particles with the surrounding plasma. Travelling wave tube amplification, Cerenkov radiation, or Doppler-shifted cyclotron generation by protons have been suggested.

The airglow is subdivided into the dayglow, twilightglow and nightglow. The sodium intensity in the nightglow and twilightglow is highest in local winter, but seasonal and diurnal variations of the other emissions are not clear as there are marked latitude effects and distinct patchiness. The origin of the nightglow is obscure, though an important part of the oxygen red emission is excited by electron-ion recombination in the F

layer. At the 85 to 100 km level there are complex chemical reactions involving oxides of nitrogen as well as the free gases and ions. The energy sources are far from understood: winds, turbulence, the quiet day ionospheric current system, thermal excitation and even particles probably contribute.

Recent research has considerably extended our knowledge of aurora and airglow. In the next few years solution of many of the problems may be expected.

References

- Chamberlain, J. W., "Physics of the Aurora and Airglow," New York and London, Academic Press, 1961.
- Störmer, C., "The Polar Aurora," Oxford, Clarendon Press, 1955.
- Sub-Committee of I.A.G.A. on Auroral Classification, "Atlas of Auroral Forms," Edinburgh University Press, 1963.

R. S. UNWIN AVOGADRO'S LAW *See* MOLE CONCEPT

B

BALLISTICS

Ballistics is an applied science dealing with the position, velocity, and acceleration of bodies in space and the forces influencing them. It is considered a branch of applied physics or MECHANICS and was formerly concerned primarily with military projectiles such as rifle bullets, field artillery shells, bombs, and rockets, but now has a bearing upon missiles, space vehicles, and orbiting satellites. Interior ballistics is concerned with the motion generated within the gun barrel or container, exterior ballistics is concerned with the motion in free air, and terminal ballistics is concerned with the effects at the point of impact (hopefully, at the target). Some bodies in motion will not involve all three divisions of this subject, for some are not launched from a gun barrel and some have no definite target or impact point (such as a vehicle which ends up orbiting the sun).

In designing a gun and a bullet, an attempt is made to obtain the highest velocity with the lowest possible maximum gas pressure. The burning rate of the propellant is important and this can be controlled by the shape and size of the pieces of propellant used. The erosion of the inside of the barrel must be held at a minimum in order to retain accuracy and keep down the heating of the barrel. It is said that some of the large German guns in World War I could be fired only once a day because they got so hot it was necessary to allow them to cool off in between firings. The smoke and the flash should be held at a minimum. Early inventors struggled with the problem of measuring projectile velocity and gas pressure in the gun barrel. The first crude method of indicating gas pressure was merely to note whether or not the gun barrel burst. When it did burst, the gas pressure had been increased too much! Great progress was made when Benjamin Robins developed the ballistic pendulum, a description of which he presented before the Royal Society of England in 1742. This involved computing the velocity of a projectile by measuring the swing of a large pendulum catching the shot. It was possible to estimate the velocity at the target as well as near the muzzle. Other methods involved the Boulengé chronograph, which measured the time elapsed between the breaking of two wire screens in the path of the shot, and the use of a magnetized projectile fired through

two coils in series which recorded the induced voltage as a function of time. Improvements came through the use of strain gauges cemented to gun barrels and x-rays to locate the position of the projectile in the barrel as a function of time, and by comparing emissivities in two spectral regions to measure or estimate the temperature of the gas.

In exterior ballistics an effort is made to trace or plot the entire trajectory of the projectile. When the target is fixed, errors of firing can be observed and corrected quickly; but when bombing is done from a plane or when the interception of a missile is attempted by launching another missile, errors of sighting or launching cannot be corrected as quickly, for the plane cannot return to its former position exactly and a second missile to be intercepted is not apt to follow the same trajectory as the first one. The first attempts to predict trajectories of shells were made on an assumption that gravity was the only force acting upon the shell, but as muzzle velocity became greater, the air resistance became many times more important than gravity. The shape of the shell became more important, and as pointed noses were designed, it became necessary to devise some way to stabilize the projectile. The shells were spun by spiral grooves inside the barrel while bombs, rockets, and missiles were equipped with fins.

As missiles became larger and more powerful they left the atmosphere of the earth sufficiently to be affected by the motion of other planets. It was difficult to record their trajectories. One method used in 1956 was to mount powerful strobe lights on the missiles and program them to flash in certain patterns. A radio signal was sent to the earth simultaneously with the flashes of the lights. Large ballistic cameras were set some on islands, along the path of the missile; and their shutters were left open during the entire flight of the missiles. The camera plates recorded the flashes of the strobe lights and also the star trails. The flights were made at night and it was important to have a clear night with no clouds. The star trails aided in the exact orientation of the cameras. An electronic computer program was devised to synchronize the timing records of the ground receiving stations and the light flashes so that the position, velocity, and acceleration of the missile were quickly computed. Computer programs were even developed to the point when the computer would predict the point of impact

from the early observations before the missile landed, usually in the ocean.

Terminal ballistics, which is concerned with fragmentation, blast, and penetration of armour and concrete, cannot be covered in detail in this brief article.

HARRY PELLE HARTKEMEIER

References

- Bliss, G. A., "Mathematics for Exterior Ballistics," New York, John Wiley & Sons, Inc., 1944.
- Davis, L., Jr., Follin, J. W., Jr., and Blitzer, L., "Exterior Ballistics of Rockets," New York, D. Van Nostrand, 1958.
- Hartkemeier, H. P., "Synchronization of Trajectory Images of Ballistic Missiles and the Timing Record of the Ground Telemetry Recording System," RCA Data Reduction Technical Report No. 31, AFMTC-TN-56-80, ASTIA Doc. No. 96634, 5 Nov. 1956, Patrick Air Force Base, Cape Canaveral, Florida.
- Hymans, J. S. C., "Guns, Shells, and Rockets: A Simple Guide to Ballistics," London, 1950.
- McShane, E. J., Kelly, J. L., and Reno, F. V., "Exterior Ballistics," Denver, 1953.
- Nelson, W. C., Ed., "Selected Topics on Ballistics," New York, Pergamon Press, 1959.
- Moulton, F. R., "Methods in Exterior Ballistics," New York, Dover Publications, 1962.
- Rosser, J. B., Newton, R. R., and Gross, G. L., "Mathematical Theory of Rocket Flight," New York, McGraw-Hill Book Co., Inc., 1947.
- Wimpress, R. N., "Internal Ballistics of Solid Fuel Rockets," New York, McGraw-Hill Book Co., Inc., 1950.

Cross-reference: DYNAMICS.

BAND SPECTROSCOPY

Theory, Instrumentation, Interpretation, and Applications. The term "band spectrum" is used to designate a spectrum originating either by emission or absorption in the molecules of a compound. As such, it is to be distinguished from a line spectrum, which is the designation of a spectrum originating in the atoms of a chemical element. It should not be assumed that the appellation band spectrum is necessarily descriptive. It is to be considered mainly of historical significance originating at a time when low-dispersion instruments were generally used for observation, which imparted an appearance of unresolved bands to many of these spectra. Actually, if an exception is made of pure rotational spectra in the far infrared region, all band spectra are highly complex when observed under high resolution. The optical radiant energy identified with molecular spectra is distributed over a very wide range of frequencies extending from the ultraviolet to the far infrared, in effect joining the microwave region at wavelengths of the order of 1 mm. The dynamic conditions accounting for this distribution will be examined briefly.

The essential differences between atomic and molecular spectra are accounted for by the fact that the former are completely explained by quantized changes in the energy associated with the outer electronic structure, whereas the latter include not only this energy, but also contributions of energy resulting from the vibration of the component atoms relative to each other and from rotation of the molecule as a whole about an axis through the center of gravity. The actual energy pattern may be represented by a summation of these three contributions; that is, we may write $E = E_e + E_v + E_r$, or following the usual notation when the energy is expressed in wave number units or term values, $T = T_e + G + F$. All three forms of energy are quantized but it may be shown that in some cases vibrational frequencies are the same as those predicted by classical mechanics. It is not necessary that all three forms of energy be present. There are instances of pure rotational energy. Since this energy is relatively small, such band spectra are found only in the far infrared beyond 20μ .

Probably the outstanding example of a completely described rotational spectrum is to be found in the paper dealing with water vapor by Randall, Dennison, Ginsberg, and Weber.¹ Vibrational spectra with rotational structure are also of frequent occurrence. Owing to the magnitudes of interatomic forces, these are found in general between 3 and 15μ .

Beginning with the observation of rotational fine structure of H_2O vapor by Eva von Bahr in 1913, and the subsequent observation of the same spectrum and those of the halogen acids by Sleator and Imes^{2,3} a short time later, the quantum interpretation of band spectra developed rapidly. The body of information on this subject is remarkably complete. It may be stated safely that all observable features of well-resolved spectra of molecules in the vapor state may be completely accounted for both within the framework of classical quantum theory and that of wave mechanics.

The most complete discussion is to be found in two books by Herzberg.^{4,5} The notation introduced in the following sections follows for the most part that employed by Herzberg. A fairly detailed treatment has been given by Ruark and Urey.⁶ Attention is also called to review articles by R. S. Mulliken⁷ and by D. M. Dennison.⁸ As the development of the subject is lengthy and detailed, only a few basic concepts are included here. In general, the theory is built up by extensions of these ideas.

In the development of the quantum theoretical explanation of the detailed features of molecular spectra, it is customary to begin with the spectra of diatomic molecules. This introduces a relatively greater degree of simplicity because one is dealing essentially with a two-body problem. Furthermore, the full details of diatomic spectra are frequently completely observable, whereas the complexity of the spectra of polyatomic molecules is in general so formidable that the

details may be resolved in relatively few instances, usually where considerations of symmetry permit a simplification of the model. The general features of polyatomic molecules may be explained by extending the concepts developed for diatomic molecules.

A physical model may be conceived for the diatomic molecule which closely represents its actual features and properties. It is thought of as a dumbbell structure. First we think of it as a rigid rotator with an axis through the center of mass perpendicular to the line joining the nuclei of the constituent atoms. A wave-mechanical treatment of this rotator leads to the following expression for the quantized energy:

$$E_{\text{rot}} = J(J+1) \frac{h^2}{8\pi^2 I}$$

where h is Planck's constant, I is the moment of inertia and J is the rotational quantum number which can take a series of integral values 0, 1, 2, This formula is generally written $F(J) = BJ(J+1)$ where the constant

$$B = \frac{h}{8\pi^2 cI}$$

Each rotational energy state is characterized by one of these J values. The quantum mechanics also leads to a requirement for a selection rule for transitions between energy states $\Delta J = \pm 1$. Since such transitions account for the emission or absorption of spectral lines the frequency of such lines is given by

$$\nu = F(J+1) - F(J)$$

or

$$\nu = 2B(J+1); J = 0, 1, 2, \dots$$

This leads to a series of equidistant lines. Such spectra are actually observed in the far infrared. The actual frequency of rotation of the rigid rotator may also be derived and turns out to be

$$\nu_{\text{rot}} = c2B\sqrt{J(J+1)} \approx c2BJ$$

That is, for any state characterized by J , the rotational frequency is approximately equal to the frequency of the spectral line that has this state as its upper state. Actually, the molecule is not strictly rigid. A correction of the simple formula is necessary to account for modification of the observed spectral features resulting from centrifugal stretching. The effect is tied in with the vibrational frequency of the molecule and leads to a modified expression for the rotation energy $F(J) = BJ(J+1) - DJ^2(J+1)^2$, where $D = (4B^3/\omega^2)$, ω being the vibrational frequency. This extra term is always small.

The dumbbell model is next considered as a harmonic oscillator, characterized by a simple harmonic motion of the constituent atoms along the line joining them under Hooke's law forces. Since the observed spectral frequency is in some instances equal to the classical vibrational frequency of a molecule of the described configuration, it is of interest to recall the formula for

simple harmonic motion

$$\nu_{\text{osc}} = \frac{1}{2\pi} \sqrt{\frac{k}{m}}$$

where k is the force constant, or force required to produce unit displacement, and m is the mass of the point moving with simple harmonic motion. For the model under consideration consisting of mass points, m_1 and m_2 , m is replaced by the reduced mass μ , equal to $m_1 m_2 / m_1 + m_2$. On the classical theory, only one vibration frequency is possible. The solution of the Schrödinger wave equation for the harmonic oscillation leads to the following equation specifying the vibrational energy states:

$$E(v) = h\nu_{\text{osc}}(v + \frac{1}{2})$$

where v is the vibrational quantum number that takes integral values 0, 1, 2, Transforming to term values the equation becomes

$$G(v) = \frac{\nu_{\text{osc}}}{c} (v + \frac{1}{2})$$

or

$$G(v) = \omega(v + \frac{1}{2})$$

where ω is the vibrational frequency in cm^{-1} . We thus obtain a series of equally spaced energy levels. The quantized emitted or absorbed energy in cm^{-1} takes place, respectively, by a transition from a higher to a lower or a lower to a higher vibrational state. The wave number characterizing such a transition is given by

$$\nu = G(v') - G(v'')$$

where v' and v'' are the vibrational quantum numbers of the upper and lower states, respectively. The selection rule that governs such transitions and which also has been accounted for by quantum mechanics is

$$\Delta v = v' - v'' = \pm 1$$

It may be noted immediately that

$$\nu = G(v+1) - G = \omega$$

indicating that the quantum mechanically derived frequency is equal to the classical frequency for all transitions. Thus we are led to expect a single vibrational absorption or emission band. Actually weak harmonics of approximately integral multiples of the fundamental frequency are sometimes observed, corresponding to transitions with Δv greater than 1. The reason for the appearance of harmonics is that there is an anharmonicity introduced into the oscillating system owing to the fact that the interatomic forces depart somewhat from Hooke's law both on close approach and extreme separation approaching dissociation. This leads to a breakdown of the selection rule. The correction for anharmonicity is in general relatively small.

The complete expression for the vibrational energy expressed as a term value is a power

series including a quadratic and for extremely precise representation of wave numbers a cubic term in powers of $(v + \frac{1}{2})$. Thus we have

$$G(v) = \omega_e(v + \frac{1}{2}) - \omega_e x_e(v + \frac{1}{2})^2 + \omega_e y_e(v + \frac{1}{2})^3 \dots$$

where the subscript e refers to the equilibrium position. Neglecting the cubic term we obtain

$$\Delta g_{v+1/2} = \omega_e - 2\omega_e x_e - 2\omega_e x_e v$$

It is convenient to measure the levels from the zero point energy obtained setting $v = 0$. In this case, $\Delta g = \omega_0 - \omega x_0 - 2\omega_0 x_0 v$.

It is prerequisite to the emission or absorption of energy that the fundamental vibration cause a change in the electric dipole moment of the molecule; otherwise it is "infrared inactive." Illustrative of such inactive vibrations are those of diatomic molecules containing identical atoms such as O_2 or N_2 . In certain instances, inactive vibrations are observable as Raman effect displacements. Infrared and Raman effect observations therefore supplement each other advantageously.

The simultaneous occurrence of vibrational and rotational transitions accounts for the observed features of infrared spectra illustrated by those of the halogen acids.

The formulation may be summarized by combining the equations already given. The term value of the vibrating rotator becomes

$$T \quad G(v) + F_r(J) = \omega_e(v + \frac{1}{2}) - \omega_e x_e(v + \frac{1}{2})^2 + \dots + B_r J(J+1) - D_r J^2(J+1)^2 + \dots$$

From this we obtain the abbreviated form of the expression for the wave numbers neglecting the correction for centrifugal stretching:

$$\nu = \nu_0 + B_r' J'(J' + 1) - B_r'' J''(J'' + 1)$$

where the values of B are appropriate to the respective vibrational states and take into account the interaction between rotation and vibration. If we set $\Delta J = +1$ and $\Delta J = -1$, respectively, and replace J'' by J , the above equation reduces to the following two:

$$\nu_R = \nu_0 + 2B_r' + (3B_r' - B_r'')J + (B_r' - B_r'')J^2$$

$$J = 0, 1, \dots$$

$$\nu_P = \nu_0 - (B_r' + B_r'')J + (B_r' - B_r'')J^2$$

$$J = 1, 2, \dots$$

These formulas represent two series of lines which are called the P and R branches. If the vibrations were strictly harmonic and if there were no centrifugal stretching accompanying the rotation or interaction between vibrations, the branches of the band would consist of equally spaced lines extending each direction from the zero energy position of the vibration band. Actually taking these effects into account the P branch shows a gradual increase in the separation of components and the R branch a decrease.

Up to this point, only cases have been considered where the electronic state of the molecule remains constant during the occurrence of vibrational and/or rotational transitions. A large class of bands, mostly in the visible and ultraviolet regions, is accounted for by a change in electronic energy, similar to that responsible for line emission spectra, accompanied by vibrational and rotational transitions. The electronic energy represents the largest contribution and accounts for the spectral location. These bands are characterized by recurrent regularities known as progressions. A rotational fine structure is superposed on the regularities. A good example is the CN bands which occur in emission with great intensity in the carbon arc. The electronic levels are characterized by vibrational structures which in turn have rotational fine structure. There is no strict selection rule for the vibrational transitions, i.e., a transition can occur between any vibrational level of one electronic state and all the vibrational levels of the other. The set of bands arising from such a group of transitions forms a progression. Similarly each of these vibrational transitions is accompanied by rotational transitions yielding the usual P and R branches and in addition a Q branch resulting from $J'' - J' = 0$. This new selection rule is accounted for by theory, based on the fact that there is a difference in the angular momentum associated with the upper and lower electronic states.

The explanation of the features of the spectra of polyatomic molecules is developed naturally from the principles outlined for diatomic molecules. Instead of a single mode of vibration, a polyatomic molecule possesses several determined by point group theory. For a molecule of N atoms there are $3N - 6$ possible modes of vibration; $3N - 5$ for a linear molecule. A fundamental radiation frequency would be expected for each of these modes. Some however are inactive, as for instance the totally symmetric vibration of a linear symmetric molecule such as CO_2 . Except for what is described as a spherical top molecule, there are always several moments of inertia corresponding to various possible axes of rotation. In general there is always a possibility of angular momentum about an axis of symmetry. This results in an extension of the selection rule for rotational quantum numbers to $\Delta J = 0$ in addition to $\Delta J = \pm 1$. There is a Q branch in these instances. An exception is the linear molecule. For example a CO_2 band shows no Q branch. As in the case of diatomic molecules, a change in dipole moment is required for emission or absorption of energy.

Even moderately complex molecules show large numbers of vibration bands, either fundamentals or in general less intense overtone or combination bands. The rotational structure is resolved only in a few instances of relatively simple molecules where the available optical resolution can be brought to bear advantageously. Examples are CO_2 , CH_4 , and NH_3 . The observed absorption bands are generally envelopes that include the

rotational energy accompanying the vibrations. Most of them occur in the "rock-salt region" between 2 and 15μ . As may be expected, some modifications in the structure of a spectrum occur when the material is in the liquid or solid state.

Molecular Spectra. The spectrum of a molecule or an atom has been described as its most definitive characteristic. The fact that the spectrum is unique leads, of course, to applications in identification and analysis that are extremely useful and important. The term "fingerprinting" of molecules is of course familiar to all analysts. The spectrum not only establishes the identification of compounds but distinguishes between various types of stereoisomerism including *cis* and *trans* forms. In order to make the best use of spectra for identification, extensive libraries of spectral absorption curves have been assembled and are being extended. Among the first to attract widespread attention was the compilation by Barnes, Gore, Liddel, and Williams.⁹ An extensive set of spectra of hydrocarbons now numbered in the thousands represents a continuing effort by the American Petroleum Institute.¹⁰ A more recent program covering a wider range of chemical compounds is that sponsored by the Coblenz Society. The material collected by the organization is published and distributed by Sadler.¹¹

The utilization of infrared analytical methods has become an important tool for industrial research, particularly in support of those industries dealing with petroleum products, plastics, and organic chemicals. The manufacture of spectrophotometric equipment is of itself an important industry. The developments in this field have been covered by biennial survey articles in *Analytical Chemistry*. The survey articles in the infrared field, including extensive compilations of references, were initiated by Barnes and Gore¹² and continued by Gore.¹³ This series provides an essentially complete index to current literature.

It was suggested by Coblenz¹⁴, more than half a century ago, that the various radicals or structural elements of a molecule should contribute characteristic features to the absorption spectrum and that it might be possible to identify compounds on the basis of a synthesis of such features. The argument has considerable validity but is not entirely conclusive. Several correlations of atomic groups and absorption frequencies in the form of charts are available.

The first was that of Barnes, Gore, Stafford and Williams¹⁵, published in 1948. The well-known Colthup¹⁶ Chart was published in 1950. "Infrared Determination of Organic Structures," by Randall, Fowler, Fuson, and Dangel,¹⁷ is largely devoted to a correlation of spectral features with various types of bonding. It also contains a compilation of spectra in chart form. The use of spectral data has made possible the identification and determination of properties of many short-lived molecules or free radicals. Examples are Ca, CH, and NH.

Many physical and chemical properties of molecules can be inferred from spectral data. The

representation of a vibration frequency by a Hooke's law formula is recalled. Knowing the masses of the constituent atoms and the frequency, it is simple to derive the force constant associated with a given type of bonding and, from this, the heat of dissociation. Conversely, from a knowledge of bond strengths, the location of the fundamental frequencies may be predicted. The rotational structure permits calculation of the moments of inertia and derivation of interatomic distances. Interesting features of geometry or spatial configuration may be established. For instance, the regular spacing of the rotational components of the CO₂ bands proves that it is a linear molecule. Similarly it has, in the instance of the water molecule, been possible to compute the angle at the O atom between the bonds connecting it to the two hydrogens.

Band spectra are important for the identification of isotopes and, hence, are useful in the detection of rare products from reactors. The isotope effect, due to the presence of Cl³⁷ along with Cl³⁵, was first observed by Imes³ and later demonstrated under high resolution by Meyer and Levin.¹⁸ The isotope effect is very large in the instance of deuterium substitution for hydrogen because of the large effect on the reduced mass of doubling the mass of one of the component atoms. It is of interest to compare the positions of the ν_2 and ν_3 fundamentals of H₂O and D₂O. The wave numbers are as follows:

	H ₂ O	D ₂ O
ν_2	1595	1179
ν_3	3756	2789

This has interesting possibilities for analytical procedures where a region of interest might be obscured by the presence of water or in certain instances its use as a solvent.

A final statement concerns the astrophysical significance of band spectra. The subject is a timely one because of the large-scale effort currently devoted to space programs, including the utilization of rocket-borne instrumentation and balloon-supported observing platforms. Since only optical radiant energy is a requirement without the necessary possession of the emitting sample, spectra may be used for a study of the atmospheres of planets or envelopes of stars. It is in this way that the presence of ammonia and methane in the atmosphere of Jupiter has been demonstrated. A detailed presentation covering the atmospheres of planets is to be found in a compilation edited by Kuiper.¹⁹ For more up-to-date information on this subject, reference is made to the proceedings of the Twelfth International Astrophysical Symposium that was held at the University of Liege, Belgium, in 1963.²⁰

CURTIS J. HUMPHREYS

References

1. Randall, H. M., Dennison, D. M., Ginsburg, N., and Weber, L. R., "The Far Infrared Spectrum of Water Vapor," *Phys. Rev.*, **52**, 160 (1937).

2. Sleator, W. W., "Absorption of Infra-red Radiation by Water-Vapour," *Astrophys. J.*, **48**, 125 (1918).
3. Ines, E. S., "Absorption Spectra of Some Diatomic Gases in the Near Infrared," *Astrophys. J.*, **50**, 251 (1919).
4. Herzberg, Gerhard F., "Molecular Spectra and Molecular Structure. I. Spectra of Diatomic Molecules," Princeton, N.J., D. Van Nostrand Co., 1950.
5. Herzberg, Gerhard F., "Molecular Spectra and Molecular Structure. II. Infrared and Raman Spectra of Polyatomic Molecules," Princeton, N.J., D. Van Nostrand Co., 1949.
6. Ruark, A. E. and Urey, H. C., "Atoms, Molecules, and Quanta," New York, McGraw-Hill Book Co., 1929.
7. Mulliken, R. S., "Interpretation of Band Spectra," *Rev. Mod. Phys.*, Parts I, IIa, IIb, **2**, 60, 506 (1930); IIc, **3**, 89 (1931); III, **4**, 1 (1932).
8. Dennison, D. M., "The Infrared Spectra of Polyatomic Molecules," *Rev. Mod. Phys.*, Part I, **3**, 280 (1931); Part II, **12**, 175 (1940).
9. Barnes, R. Bowling, Liddel, Urner, and Williams, Van Zandt, "Infrared Spectroscopy. Industrial Applications," *Ind. Eng. Chem., Anal. Ed.*, **15**, 659 (1959).
10. American Petroleum Institute Project 44, Petroleum Research Laboratory, Carnegie Institute of Technology, Pittsburgh, Pa.
11. Sadtler Research Laboratories, 1517 Vine Street, Philadelphia 2, Pa.
12. Barnes, R. B., and Gore, R. C., "Review of Fundamental Developments in Analysis—Infrared Spectroscopy," *Anal. Chem.* **21**, 7 (1949).
13. Gore, R. C., *Anal. Chem.*, **22**, 7 (1950); **23**, 7 (1951); **24**, 8 (1952); **26**, 11 (1954); **28**, 577 (1956); **30**, 570 (1958).
14. Coblenz, W. W., "Investigations of Infrared Spectra," *Carnegie Inst. Wash. Publ.*, **35** (1905). Reprinted by the Coblenz Society and the Perkin-Elmer Corporation, 1962.
15. Barnes, R. Bowling, Gore, R. C., Stafford, R. W., and Williams, V. Z., "Qualitative Organic Analysis and Infrared Spectrometry," *Anal. Chem.*, **20**, 402 (1948).
16. Colthup, N. B., "Spectra-Structure Correlations in the Infrared Region," *J. Opt. Soc. Am.*, **40**, 397 (1950).
17. Randall, H. M., Fowler, R. G., Fuson, Nelson, and Dangi, J. R., "Infrared Determination of Organic Structures," Princeton, N.J., D. Van Nostrand Co., 1949.
18. Meyer, C. F., and Levin, A. A., "Absorption Spectrum of Hydrogen Chloride," *Phys. Rev.*, **34**, 44 (1929).
19. Kuiper, G. P., "The Atmospheres of the Earth and Planets," Second edition, Chicago, The University of Chicago Press, 1952.
20. "Les Spectres Infrarouges des Astres (Region: 1μ à 3 mm), *Extrait Mém. 8^e Soc. Roy. Sci. Liège, Cinquième Sér.*, **9** (1964).

Cross-references: ABSORPTION SPECTRA, ATOMIC SPECTRA, SPECTROSCOPY.

BATTERIES

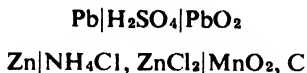
The term "battery" generally denotes an assembly of one or more electrochemical cells which exhibit a voltage difference between their two electrodes, so that they will deliver usable current when connected to an external circuit. Some authorities apply the term "galvanic" or "voltaic" cells to this family, to distinguish them from cells to which current is fed from an outside source for purposes such as electroplating or extraction of metals. There are many ways in which a voltage difference can be produced in an electrochemical cell. The simplest cell, thermodynamically speaking, is the "concentration cell" in which electrolyte or electrode materials are incorporated into half-cells in differing concentrations; a "half cell" is a system involving an electrolyte and a single electrode. When the half-cells are connected, the free energy change accompanying the transfer of one substance from high to low concentration results in the liberation of electrical energy. Concentration cells, though interesting theoretically, are not important commercially.

The majority of economically important cells consists of two dissimilar electrodes of metal or metal compounds, immersed in an aqueous solution of an acid, base, or in some cases a salt. The negative of a fresh cell is typically in the metallic state, while the positive is usually an oxide, or occasionally, a salt of the metal. During discharge, the negative electrode is oxidized as electrons leave it via the external circuit, and the positive is reduced. Since by definition an anode is an oxidation electrode, in the literature the negative is generally called the "anode" and the positive the "cathode"; this conforms to accepted electrochemical terminology, though it is the cause of some confusion.

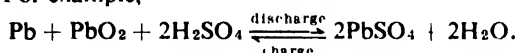
Theoretically, galvanic cells might look more attractive than heat engines as sources of electric power, since the energy changes are not subject to the limitations of the Carnot cycle. In actuality, however, the cost of public-utility-produced power is 3¢ to 5¢ per kWh, as compared with roughly \$2 per kWh for the power delivered throughout the life of a sealed rechargeable nickel-cadmium cell, and \$50 per kWh for a non-rechargeable Leclanché flashlight cell. This is because of inefficiencies in electrochemical operation, high material costs, high cost of the tightly controlled production operations necessary, etc. Galvanic cells have grown in importance because of the strength of other needs such as the need for a portable supply of power, for power at a place far distant from the prime power source or at a different time than can be supplied by the prime source, for a reserve supply of power to cover peak demands which are beyond the capacity of the prime source or to supply power in the event of failure of the prime source. There are also needs for a source of pure direct current or for a stable reference voltage which are filled by galvanic cells.

Taking the familiar lead-acid storage battery and the conventional Leclanché flashlight-type

cell as examples, cell systems are schematized in the literature as follows, the vertical lines denoting phase boundaries between the solid active materials and the electrolyte:



Abbreviations denoting the state and the concentration of the materials may appear beside the chemical symbols. Some cells are shown as containing more than three phases; e.g., those with different electrolyte at each electrode. In all cases, the oxidation electrode, the negative, is given on the left, and the reduction electrode on the right, with the electrolyte phase or phases between. Cell reactions are written in the conventional chemical fashion, the charged reactants on the left and the discharged products on the right. For example,



References which discuss cell reactions in any detail will include also the half-cell reactions for each individual electrode.

The basic thermodynamics of a galvanic cell operating reversibly at constant temperature and pressure is described by the equation

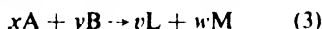
$$\Delta G = -nF\xi \quad (1)$$

where ΔG is the change in the Gibbs free energy, n is the valence change in the cell reaction, F is the Faraday (96,487 coulombs), and ξ is the reversible emf of the cell. In terms of ΔH , the heat of the reaction,

$$\xi = -\frac{\Delta H}{nF} + T\left(\frac{\partial \xi}{\partial T}\right)_p \quad (2)$$

where T is the absolute temperature, and the derivative is the voltage-temperature coefficient of the cell at constant pressure.

Expressing the basic cell reaction as



a practical expression for ξ may be written:

$$\xi = \xi^0 - \frac{RT}{nF} \ln \frac{(\alpha_{\text{L}})^v (\alpha_{\text{M}})^w}{(\alpha_{\text{A}})^x (\alpha_{\text{B}})^y} \quad (4)$$

where R is the gas constant, α_z the activity of species Z and ξ^0 is the standard emf; the latter term, ξ^0 , may be expressed as

$$\xi^0 = \frac{RT}{nF} \log K \quad (5)$$

where K is the equilibrium constant of the cell reaction. Alternatively, ξ^0 may be calculated in terms of the oxidation potentials of the negative and positive electrodes; the potentials of many electrode reactions may be obtained from standard sources.¹ A rough calculation of ξ may be made by using the concentrations of the several species as approximations of their activities; the

activity of a pure phase, such as a solid, is considered to be unity.

In actual operation, a discharging cell does not behave reversibly, and therefore does not obey Eq. (1). Hence it is useful to define the electrochemical efficiency ϵ as the ratio of the chemical energy expended to the electrical energy delivered during the time t_D necessary to completely discharge the cell; i.e.,

$$\epsilon = - \int_0^{t_D} \xi_D i_D dt / \Delta G \quad (6)$$

where ξ_D in this case is the cell terminal voltage, and i_D the current, during discharge. With rechargeable cells, it is also useful to define the storage efficiency ϵ_s as the ratio of the energy delivered during discharge to that returned in the subsequent recharge (denoted by subscript C); i.e.,

$$\epsilon_s = \int_0^{t_D} \xi_D i_D dt / \int_0^{t_C} \xi_C i_C dt \quad (7)$$

In the definition of ϵ_s , it is assumed that the cell is charged until it is returned to its original condition.

Both ϵ and ϵ_s are considerably less than unity in practical operating cells, for a number of reasons. Not all the current that theoretically could be delivered by the quantity of active material in the cell will actually be produced, due to incomplete chemical reactivity, imperfect electrical contact between the active material and the conductive grid, and other factors. Further, observed cell voltage will be less than that predicted from thermodynamic data, due to competing reactions occurring spontaneously within the cell and to "polarization" effects (polarization is the term given to any process, such as the accumulation of products of the discharge reaction at the reaction sites, which causes a reduction in the potential of an electrode as current is being passed through it).

Typically, ϵ will range from 75 to 90 per cent (the latter figure pertaining to alkaline cells with low internal resistance, operating at low discharge rates); ϵ_s will generally range from 55 to 75 per cent, depending, as one factor, on the rate at which charging is done and the consequent spread between charge and discharge voltage. With practical cells, the degree to which the change in free energy of the cell reactants can be realized in terms of electrical energy is not of as much importance as is the electrical energy produced per unit weight or unit volume of the cell.

For the sake of providing some comparison between galvanic cells and thermal converters, consideration might be given to thermal efficiency. This can be approximated by substituting ΔH for ΔG in Eq. (6). Thermal efficiency in many modern cells is 50 per cent to 80 per cent, as compared with 40 per cent to 45 per cent for the better thermal converters. Thermal converters may, however, be superior on a watt-hour per pound basis, particularly when operated over long periods of time.

Physical Configurations, Plate Types. Cells are commonly made in three basic configurations: cylindrical, rectangular, (sometimes called "prismatic," particularly in Europe), and small disk-shaped "button" cells. Each of these has distinct advantages and disadvantages, and each is made with plates of specialized design, tailored to fit that shape. Cylindrical cells make optimum use of internal volume and contain internal pressures well (a necessity with sealed cells). When assembled into multicell batteries, however, they are wasteful of battery internal volume. Further, they require specialized, and often costly, production techniques. Plates for cylindrical cells are frequently molded as slugs of the desired cylindrical shape, or they may be cut from sheet material into long strips which then are rolled, with an intervening strip of separator, into a coil. This sheet plate material may be made by a wide variety of methods including sintering, electroplating, pasting of active materials onto some suitable support, impregnating active materials into a porous matrix, or forming sheet metal. In rare cases, flat rectangular plates may be assembled into a cylindrical cell container.

Rectangular cells are easily assembled; the cases are easily fabricated of a variety of materials. They make acceptable use of battery internal volume. They do not, however, contain well the force of plate expansion or internal gas pressure. And usually a lesser proportion of cell internal volume can be filled with plate material than can be done with cylindrical cells. Button cells lend themselves to automatic production techniques; the assembly operations are relatively simple and inexpensive; the cells are easily stacked into batteries of any required voltage. As disadvantages, they do not contain internal pressures well; they cannot easily be made in sizes much larger than 1.0 ampere hour in capacity; and they involve considerable waste of cell internal volume. Plates for both rectangular and button cells are almost invariably flat. Button cell plates are usually molded or stamped from sheet stock into the required disk shape. Plates for rectangular cells may be sheet material made by any of the methods given in the previous paragraph. Or they may be cast or molded.

"Primary" and "Secondary" Types. Galvanic cells of commercial importance fall into two classes, "primary" and "secondary." Primary cells are said to be nonreversible: at the end of discharge, the active materials are in such state that they cannot be returned to their original condition by charging. Secondary cells are said to be reversible; they can be recharged by applying current to the cell terminals in opposite polarity to that produced by the cell during discharge. As normally used, these cells are not truly reversible in the thermodynamic sense: keeping charge-discharge times within acceptable limits, it takes considerably more than an "infinitesimal increase" in opposing force to drive the cell reaction in the opposite "charging" direction. Ordinarily, primary cells yield higher energy densities (ratios of power output to weight or volume) than

do secondary cells. One reason for this is that providing rechargeability involves considerable sacrifice in the amount of reactive material that can be packed into a given unit weight or volume, due to the need to incorporate relatively massive grid structures within or around the plates, plus space- and weight-consuming separators between the plates. Providing rechargeability may even require incorporating the active materials in a form which offers less than optimum electrochemical reactivity.

Primary Cells. Until recent years, zinc was used as the anode of almost every primary cell known. The metals above zinc in the electromotive series (aluminum, magnesium, calcium, potassium, etc.) react so readily with the water of the aqueous electrolytes (the only electrolytes judged practical in the past), that much of their energy is consumed in wasteful side reactions. The metals below zinc provide inadequate cell voltage and, in several instances, are costly (chromium, cadmium, cobalt). Zinc is still the only common primary anode, although magnesium, calcium, lead, silver, and indium have been used for some special military and experimental cell types. A considerably greater variety of materials has been used as cathodes, including, in the older cells, manganese dioxide, oxygen, cuprous oxide and copper, carbon, and mercuric oxide. More recently, primary cells have been built with cathodes of mercuric oxide, silver (I) and silver (II) oxide, lead dioxide, zinc chloride, lead chromate, copper halides, silver chloride, and, in experimental cells, such materials as bismuth trioxide, vanadium pentoxide, and a variety of organic compounds such as the nitrobenzenes, nitropropanes, and N-halogens which promise high energy yield per unit weight.

Secondary Cells. Far fewer secondary than primary systems have been developed to a practical level. The difficulties in providing acceptable cost, satisfactory physical, electrical, and environmental characteristics, together with rechargeability, are so severe that today there are only six secondary systems of any practical importance. Five of these—the nickel-cadmium, nickel-iron (Edison), silver-zinc, silver-cadmium, and zinc-manganese dioxide—are alkaline. (In Europe, particularly the Soviet Union, another alkaline system, the nickel-zinc, is used.) The remaining system, the familiar lead-acid, is the cheapest storage battery type, and is by far the most important commercially. Total sales of all secondary types in 1962 were reported to be \$460 million; of this, \$425 million was for lead-acid batteries, the preponderant majority of which were ordinary car or truck starting batteries.

Standard Cells. Standard cells are galvanic cells that are used to provide a stable, known voltage for calibration of other equipment, rather than for supplying electric power. The only type in general use now is the "Weston," or mercury-cadmium cell. The "saturated" cell type is: $\text{Cd} | \text{Hg}_3\text{CdSO}_4 \cdot 8\text{H}_2\text{O} | \text{CdSO}_4 | \text{Hg}_2\text{SO}_4 | \text{Hg}$. The anode is a 10 per cent cadmium amalgam; the

CdSO_4 is a saturated aqueous solution. The emf at 20°C is 1.01864 volts. It has a fairly large temperature coefficient (0.00004 volt/ $^\circ\text{C}$ at 20°C). A more generally used type, the "unsaturated" cell (assembled without cadmium sulfate crystals) has a temperature coefficient of 0.00001 volt/ $^\circ\text{C}$.

Fuel Cells. Fuel cells, though operating as galvanic cells at the electrodes, are in a class by themselves in that they provide direct, single-site conversion of original raw materials into electrical power, obviating the boiler-turbine-transmission-rectifier chain that precedes the production and use of ordinary batteries. The raw materials, usually involving hydrogen and oxygen, are fed into the cell in a constant flow, as long as power is needed; when the fuel flow is stopped, the cell is inert. The fuels react at the electrodes, forming H_2O as the exhaust in most cells, and liberating electricity. Calculating actual power output versus theoretical free energy change of the original fuels, efficiencies of 65 to 85 per cent have been achieved in prototype and laboratory models. The technological problems remaining to be solved before practical fuel cells appear on the market are formidable, and include elimination of the water from the pores of the electrodes, preventing contamination of the electrochemical system by impurities carried in with the cheaper fuels, achieving workable energy densities, simplifying the necessary control and supporting equipment, and providing adequate electrode life. For an excellent general discussion of various fuel cell types, see reference 2.

WM. W. JAKOBI

References

1. Latimer, W., "The Oxidation States of the Elements and Their Potentials in Aqueous Solutions," second edition, New York, Prentice-Hall, Inc., 1952.
2. Mitchell, Will, Jr., Ed., "Fuel Cells," New York, Academic Press, 1963.
3. "Batteries and Electric Cells" in "Encyclopedia of Chemical Technology," second edition, Vol. 3, New York, John Wiley & Sons, Inc., 1964.

Cross-references: ELECTRIC POWER GENERATION, ELECTROCHEMISTRY, POTENTIAL.

BETATRON

The betatron is a particle accelerator using a sustained induced voltage to accelerate the particles to full energy during the whole period of acceleration of the particle. Since this method of acceleration seemed most applicable to electrons, the name "betatron" was used to indicate that it was the agency for producing high-speed electrons. The action of the betatron is similar to the action of an electrical TRANSFORMER in which a high-voltage winding of many turns is used. In a transformer, the voltage can be stepped up from the primary voltage, V_1 , to the secondary voltage, V_2 .

$$V_2 = V_1 N_2 / N_1$$

where N_2 is the large number of turns of the secondary and N_1 is the small number of turns in the primary. For example, x-ray transformers having high voltage, such as 100,000 volts, have very many turns of fine wires and consequently raise the primary voltage by a large factor.

The main structure of the betatron is really a transformer, and focusing magnets are arranged around the transformer core where a secondary winding might be put. A vacuum tube instead is placed between these focusing magnets so that it can conduct electrons hundreds of thousands of times around the core. Each time the electron circulates around the core it acquires an energy equivalent to the voltage which would have been induced in one turn of wire at that instant.

In order to guide the electrons on many revolutions the focusing magnet can be such that the magnetic field decreases with increasing radius. Then the lines of force bulge outwardly about the orbit, providing vertical focusing forces back to a horizontal orbit in case the particle strays above or below the orbital plane. This bulging field is the requirement for vertical focusing, and it is used in cyclotrons too. However, it is necessary that the magnetic field decrease less rapidly than $1/r$, where r is the radius of the orbit. If this latter requirement is met, then radial focusing will be insured; because the required centripetal force to hold the particle in the circle going around the core decreases as $1/r$. Consequently, if the magnetic force decreases less rapidly than this, it will be too strong at large radii to permit the orbit to remain circulating at a large radius, and it will be too weak at small radii to maintain the particle circulating at a small radius. The particle thus will hunt about the so-called equilibrium orbit by means of this radial focusing action in addition to the previously mentioned vertical focusing action. The oscillations of the particle about this orbit are called betatron oscillations, and the name appears in the scientific literature referring to particle motions in other accelerators such as synchrotrons, because this focusing was first worked out for the case of the betatron.

It is possible to have strong focusing magnets with much more rapid variation of field with radius to provide the focusing. In this case, a succession of focusing and de-focusing magnets must be used which alternately focus vertical and radial motion and which are called alternate gradient focusing magnets. Such strong focusing magnets can limit the oscillation of the particle about the equilibrium orbit to a very small amplitude.

The usual betatron has a magnetic field which rises proportionately with the increase in the transformer's magnetic flux within the orbit. By means of this proportionality, the strength of the guiding field rises along with momentum gained by transformer action, and hence the guiding field provides sufficient centripetal force to hold the particle at the same radius.

The first betatron of this type produced 2 MeV and radiation equivalent to 2 grams of radium.

It is now in the Smithsonian Museum in Washington, D.C. This accelerator is the size of a typewriter. The largest betatron can generate beams of 320 MeV. The x-rays and electrons can be used to produce mesons and numerous nuclear disintegrations. The most commonly used betatrons are for 25 to 35 MeV. These provide x-rays of maximum penetration in iron for industrial radiography, and they provide x-rays and electrons with optimum depth dose characteristics for x-ray or electron therapy of the human body.

The intensity of radiation from the 25-MeV betatron is of the order of 100 to 200 roentgen/min at a meter from the target for x-rays. With the extracted electron beam, the ionization doses would depend on how widely the electrons are spread at the point of treatment, but comparable doses are obtained. The large 320-MeV betatron at the University of Illinois produces intensities of the order of 20 000 roentgen/min at a meter or, in other terms, of the order of 5 watts.

It is possible to make what are called fixed-field alternating-gradient betatrons (FFAG). In this case, the focusing field is constant in time, and the particle orbit can be caused to spiral either outwardly or inwardly with increasing energy between the focusing poles of a direct current magnet or a permanent magnet. These FFAG betatrons are capable of giving a beam of x-rays up to 20 per cent of the time, whereas the betatrons with time-varying focusing fields just give one pulse of electrons in every cycle, and these pulses are only of the order of a microsecond in duration. Because of the large duty factor available with FFAG betatrons, it should be possible to achieve intensities of 10,000 watts. Although FFAG combination betatrons and synchrotrons have been made, full advantage has not yet been taken of a large duty factor achievable by incorporating a full-size transformer core within the betatron orbit.

In the case of the conventional betatron with constant orbit radius, the relation between the strength of guiding field and the total flux change within the orbit can be found as follows:

The momentum of the orbit

$$p = \frac{e}{c} BR$$

while the rate of change of momentum

$$\frac{dp}{dt} = eE$$

where e is the charge of the electron, c is the velocity of light, B is the magnetic field in gauss, R is the radius in centimeters, and E is the electrical field in electrostatic units.

$$2\pi RE = \frac{1}{c} \frac{d\Phi}{dt}$$

the volts per turn where Φ is the flux linking the orbit. Combining the last two equations

$$\frac{dp}{dt} = \frac{e}{2\pi Rc} \frac{d\Phi}{dt}$$

Thus, after a lapse of time

$$p = \frac{e}{2\pi Rc} (\Phi_2 - \Phi_1)$$

Combining this with the first equation

$$2\pi R^2 B = \Phi_2 - \Phi_1$$

Thus the flux change within orbit $\Phi_2 - \Phi_1$ is twice as big as the flux would be if the flux density, B , were constant within the orbit. Therefore, the transformer core must be adjusted so that the proper excess flux density provided within the orbit meets the conditions of this last relation if the orbit is to be at a constant radius, the assumption made in the above derivation.

D. W. KERST

Cross-references: ACCELERATORS, LINEAR; ACCELERATORS, PARTICLE; CYCLOTRON; SYNCHROTRON; ACCELERATORS, VAN DE GRAAFF.

BIONICS

Bionics is defined as the branch of knowledge pertaining to the functioning of living systems and the development of non-living systems which function in a manner characteristic of, or resembling, living systems. The definition of bionics infers the use of scientific skills and techniques from biological, physical, mathematical, and applied sciences in carrying out research in which: (1) the functions of chosen biological components and systems are studied and analyzed to determine underlying principles and processes that may lead to methods for improving physical components or systems, and (2) the theories and techniques of chemistry, physics, and mathematics are applied to advance our knowledge of the principles upon which these functions are based.

Bionics research depends on the acceptance of certain postulates. These postulates are of two types—one essentially operational, the other essentially technical. The important operational postulates are: (1) common experience shows that biological systems, for example the human nervous system and its peripheral equipment, perform operations such as pattern recognition and identification, discrimination, and learning that no non-biological system can now perform efficiently; (2) other biological components perform such functions as detection, filtering, and information transfer more efficiently and with greater certainty over broader bandwidths than do present non-biological components; (3) intensive study, analysis, and application of the principles that make superior biological performance possible can lead to non-biological systems that equal or may, in some cases, exceed biological systems capabilities. The technical postulates are: (1) The functional advantages of biological systems are implied in the unique methods for information transfer, memory storage, and retrieval, united with unique ways for correlating and integrating data from many

sensors or sensor systems. These unique methods depend mainly on: (a) many converging and diverging information transfer channels and many connections between channels, and (b) the special properties of biological components at the points where these channels are interconnected. (2) The superior capabilities of biological components or particular elements of the components (for example, receptor cells or nerve muscle junctions) are derived from specific ways of interconnection and probably as well from specific molecular properties. (3) Analysis and study of the relevant data describing biological systems that will represent major improvements over existing physical and engineering hardware. (4) The present rapid advances in microminiaturization techniques suggest that, for the first time, the possibility exists for the development of physical components or systems that incorporate these superior biological principles and processes.

The definition of Bionics suggests the methodology of procedure. We design, grow, or in some way obtain non-biological systems that function in a way "resembling" living systems. The physical component simulates the biological way of doing or carrying out its function. To obtain this objective, we first choose the biological components that perform the desired function; second, we compile the descriptive biological data; third, we translate this data into engineering terms; and fourth, we apply the translated data for the physical simulation of the function.

This process requires application of relevant mathematics to describe clearly, and as rigorously as possible, the biological function by some mathematical theory or model. This process may also require various techniques from the physical sciences to arrive at the necessary data defining the biological function of interest.

While the bionics research procedure requires the description, mathematically, of the function to be performed by the non-biological system, it may also require mathematical and physical descriptions of the properties of the materials used to construct the non-biological analog.

Gaining enough data to describe the given biological function may require study of the biological component or one of its elements at the molecular level. Similarly, solid-state or molecular techniques may be required to construct the physical system that is the appropriate simulation of the desired biological function.

Successful bionics activity may be modified by several factors. These include the complexity of the biological functions; the kind, quality, and quantity of data available to describe the functions; and the existence of relevant mathematical and physical techniques essential to the simulation of the functions.

In terms of the bionics objectives, the scope of the work must include research on the following components and/or systems:

- (a) receptors or sensors;
- (b) receptor systems—including central interconnections and interconnections among receptor systems;

- (c) central nervous system networks and the interconnections among parts of the nerve network;

- (d) effectors and actuators;

- (e) effector systems—including the feed-back and feed-forward connectors to and from the central processor;

- (f) the integrated system made up of sensors and their input channels, the central correlating, control, and computing networks, and the channels from the central system to the effectors and actuators.

This complete research program will extend over a long period of time. However, progress has been made. Data have been acquired, analyzed, and put into engineering terms to provide a set of specifications for the functional properties of several types of neurons. The neuron is presumed to be that physiological component underlying observed psychological parameters such as learning, adaptation, etc. Therefore, the construction of a network of artificial neurons should, to some extent, simulate these observed behavioral parameters. In addition, data have been acquired and analyzed from the field of experimental psychology on functions such as learning. These data can be translated into engineering and can suggest types of components which can simulate the learning function directly. This example illustrates a procedure which is common in bionics research. One can start at the operational level with an observed set of functional parameters and attempt to synthesize the class of mechanisms that could simulate these functions, or one can start with the biological component which is presumed to give rise to these functions and attempt to simulate this component. The choice of one method over the other depends greatly on the assurance one has of the validity of the data at one level or the other.

CECIL W. GWINN
LEONARD M. BUTSCH, JR.

References

- "Bionics Symposium 1960," Wright Air Development Division, U.S. Air Force, Technical Documentary Report 60-600.
- "Biological Prototypes and Synthetic Systems," New York, Plenum Press, 1962.
- "Bionics Symposium 1963," Aeronautical Systems Division, U.S. Air Force, Technical Documentary Report 63-946.
- Thompson, D'Arcy W., "On Growth and Form," Cambridge, Cambridge University Press, 1942.
- Defares, J. G., and Loneddon, I. N., "Mathematics of Medicine and Biology," Chicago, Year Book Medical Publishers, 1961.
- Rashevsky, N., "Mathematical Biophysics," Third edition, New York, N.Y., Dover, 1960.
- Sommerhoff, G., "Analytical Biology," London, Oxford University Press, 1950.
- Ashby, W. R., "Design for a Brain," New York, John Wiley & Sons, Inc., 1960.
- Ashby, W. R., "Introduction to Cybernetics," New York, John Wiley & Sons, Inc., 1958.

- Elasser, W. M., "Physical Foundations of Biology," New York, Pergamon Press, 1958.
- Ashby, W. R., "Homeostasis," in Gray, P., Ed., "Encyclopedia of Biological Sciences," New York, Reinhold Publishing Corp., 1961.
- Steele, J. E., "Bionics," in Susskind, C., Ed., "Encyclopedia of Electronics," New York, Reinhold Publishing Corp., 1962.

Cross-references: BIOPHYSICS, CYBERNETICS, MATHEMATICAL BIOPHYSICS.

BIOPHYSICS

Just as biochemistry was born slowly and painfully out of general physiology over a period of fifty years centered at the turn of the century, so biophysics is now emerging as a major, distinguishable part of science. Disciples of this youngest daughter of general physiology, biophysicists apply the new and developing techniques of physics and physical chemistry to biological subject matter; in addition, they develop new methods to examine, and theories to describe, biological processes.

Because of its biological and physical origins, the subject cuts into, or overlaps with, several other "branches" of science. However, like all these other disciplines, the boundaries are not too well defined, nor can they be in a classification which is essentially a subjective one. The various subdivisions of biophysics—mathematical, physical, physicochemical, physiological and psychological—intrude on and overlap with the corresponding established disciplines whenever those disciplines make a physically oriented attack aimed at the fundamental description and understanding of living matter.

Certain men in history focused the experience of their time. Biophysics might be said to have begun immediately after Isaac Newton was hit on the head with the falling apple: effects of gravitational attraction and the psychological results of the impulse to his head doubtless drew his immediate attention! Interest in impulse and momentum-transfer to living matter persists for obvious pathological reasons.

The discovery of animal electricity by Luigi Galvani in 1778 opened up, in Alessandro Volta's words 20 years later, "a very wide field for reflection." The development of electrophysiology and the more fundamental biophysical studies of the molecular and ionic mechanisms associated with irritability of tissue are a distant consequence.

The anatomist, Adolph Fick, considered by some to be the father of biophysics, showed the way to quantitative studies of diffusion and mass transport. Recognizing immediately the interdisciplinary nature of his studies, he thoughtfully published his work simultaneously in a journal of physics and a journal of physiology in 1855. Modern familiarity with exchange processes, membrane processes and the like, stems from his pioneering work. Meanwhile Thomas Young was completing a long career during which he

formulated important ideas on the physics of color vision and heart action. Young was followed later into studies on these and many other topics of biophysical interest by one of the intellectual giants of the nineteenth century, the ubiquitous scientific explorer, Hermann Ludwig Ferdinand von Helmholtz, who put color vision, hearing, and energy conversion, on a physical foundation which was firm enough to withstand nearly a century's further progress.

The basic principles of fluid flow through blood vessels, and pipes in general, were described quantitatively in 1884 by Jean Louis Poiseuille. It was only a few years later that that last fruitful quarter-century was completed in rousing fashion by three other discoveries which were to lead to later very important developments in biophysics. First, Svante Arrhenius, Hans Kohrausch, and others characterized the electrical and osmotic properties of *ionic* solutions. Second, Henri Becquerel discovered natural radioactivity in 1895; and third, Wilhelm Roentgen discovered the penetrating and ionizing radiation which he named x-rays in 1898. From the new fundamental knowledge of electrolytes, have proceeded electrochemical studies on the origin and nature of electric currents in living tissue, and studies on the subtleties of water balance in cells and tissues. In a similar development, further knowledge of the origin, physical properties and the physical effects of particulate and electromagnetic ionizing radiations has formed the basis for studies of their biophysical effects, which are closely interwoven with those being studied in radiation chemistry, radiobiology and radiology. For example, studies on the time and spatial distribution of the energy absorbed by living tissues from gamma-ray beams have led to some understanding of the mechanisms of (a) changes induced in mass-transport processes in blood plasma, (b) changes in chemical reactivity, and (c) subtle changes in the electrochemical processes which form the basis of transmission of information along nerve membranes. The psychological effects of an increased background of ionizing radiation on the noise level in human nerve transmission may be one of man's important problems in the future.

The continuing development of fast, sensitive electronic equipment—fast dc amplifiers, electron microscopes of high resolving power, ultraviolet microscopes, mechano-electrical transducers, etc.—has encouraged more and more physicists to become interested in the properties of biological material. Conversely, students of biology have become increasingly more familiar with properties of fundamental particles, physical measurements, kinetic techniques and analyses, etc., and are applying them to their focus of interest when feasible. For instance, during the 1930's some very clever experimental and theoretical studies were begun by A. V. Hill and others on the transformation of the energy from chemical reactions into the mechanical energy of muscle contraction and the attendant electrochemical energy for nervous control. Although

the process is unbelievably complicated on the molecular scale, research is now able to be centered on myosin, the remarkable contractile protein-enzyme molecule, and its molecular associate, actin. Chemical activity of myosin as a specific hydrolytic enzyme is accompanied by conformational changes, which permit both sliding and folding processes to occur at the macromolecular level during contraction of the bulk muscle. The energy balances (heat, mechanical, electrochemical, chemical) and the power balances are slowly being worked out in detail. They are devious, reflecting the complexity of the basic molecular system. Although still only partly quantitative, a comprehensive molecular theory is becoming more and more satisfying.

Since the early 1950's, a concerted study of the interaction of matter waves (sound and ultrasound) with living tissues has been apparent, with contributions from several disciplines: electrical engineering studies on the production, focusing and measurement of radiated sonic energy; physiological studies on the reception by the ear and other mechano-receptors; electrochemical studies of interfaces irradiated with acoustic beams, associated with the receptor-nerve junction; physical studies of the mechanisms of absorption by watery polymeric solutions. From these has emerged a flourishing part of biophysics not only concerned with the fundamental processes of absorption and transduction, but also providing knowledge useful in the diagnosis and location of brain disorders and subcutaneous lesions, and even therapy by acoustic irradiation.

A mathematical description of selected biophysical concepts was made in 1938 by Nicholas Rashevsky, and in the years since, he and one group of biophysicists have been actively developing mathematical descriptions of current ideas on growth, cancer, energy and mass transfer in cell respiration, cell movements and, more recently, nerve conduction and the stability (steady-state) of living organisms (see MATHEMATICAL BIOPHYSICS for Rashevsky's account of this research). Inherent in this work are developments of new mathematical concepts, or algebraic models, which at least in principle can describe the detailed behavior of living tissue when stimulated with certain forces. Perhaps the greatest contribution of this kind of work is that it presses the experimental biophysicist continually to refine his techniques, to measure yet more parameters of the living material he is studying, for the models developed usually contain explicit functions of parameters which are not or cannot yet be measured. The difficulty in this work is to be able to relate the theory, at sufficiently small intervals in its development, to reality as found by the experimentalist.

The quantitative concepts so successfully used in the electrical and electronic engineering of guided missiles have recently been applied with useful results to processes occurring in the brain, and to the better understanding of the principles of organization of the physical apparatus which

forms the physical basis of memory, emotions, and instinct. Even without complete knowledge about how a neuron works, in terms of atoms and molecules of which it is composed, certain properties of cell assemblies, or neural nets, are beginning to show in research on electronic models, and the systems concept is suggesting where else we should look for other properties of other subsystems, or assemblies. This biophysical approach may, in another twenty years, enable man to give a better physiological as well as psychological account of neuroses and anxieties. Maintenance of the steady-state, assessment of the kind of and time of response to given inputs or feedbacks—in short, problems in stability of the system—occupy an important group of biophysicists today. Even the teaching of biophysics has received benefit from the early introduction of the systems concept into the course: it is a good framework on which the teacher can hang the details of the course.

Meanwhile a considerable number of biophysicists concern themselves with two very new and modern topics: molecular biophysics and the biophysics of weightlessness. This work seems to have to be done by big teams, in which the biophysicist has his role. Evidence was presented in 1961 that a coding of genetic and hereditary information in the chromosome may be carried in the arrangement of the four pyridine and pyrimidine bases which occur in desoxyribonucleic acid (DNA), and since then other macromolecules have become suspect. Further on macromolecules: for many years certain vague variants of composition and structure had been associated with specific so-called molecular diseases. As a result of these two developments, the subject of "molecular biology" has become loosely circumscribed as a special very important interest in science; and as part of this circumscribed interest has sprung "molecular biophysics"—loosely (yet) defined as the study of structure and physical properties of macromolecules as they affect their physicochemical roles in living tissue.

With man now entering the space age, the problems he and his living ancillary equipment will have to face are already becoming apparent. Lessening of elasticity in the major arteries of the astronauts—an accommodation to less stress on the walls in the absence of gravitational force—has recently been proven, and its mechanism is being sought. Plant cells fixed and stained in space showed distortions of the cell membranes, as well as disorder in the normally well-organized nucleus, mitochondria, and other cell components. The absence of gravitational force apparently has new and unpredictable effects on heretofore earthbound biological material. It probably should be mentioned in passing that we can only guess now about what changes will occur in living tissues as they leave the electromagnetic field of the earth with which so many biological rhythms seem to be bound.

In the words of A. V. Hill, biophysics is the study of biological function, organization and

structure by physical and physicochemical ideas and methods. The application of techniques and fundamental principles in studies of living tissues and the recognition of new principles and the development of new methods from them comprise biophysics.

An International Union of Pure and Applied Biophysics recognizes the scope of the subject, and many countries have national Biophysical Societies. Learned journals devoted to research, and review papers in biophysics include: *Advances in Biological and Medical Physics*, *Progress in Biophysics and Biophysical Chemistry*, *Biophysical Journal*, *Acta Biophysica*, *Physics in Medicine and Biology*, and *Archives in Biochemistry and Biophysics*. The *Proceedings of the Institute of Electrical and Electronic Engineering* has had notable issues on biophysical topics. Among the contributed volumes designed to broaden the view of the specialist is "Biophysical Science—A Study Program," edited by J. L. Oncley *et al.*, John Wiley & Sons, Inc., New York, 1959.

Textbooks at various levels of study, and reflecting the individual authors' experience and viewpoints of this broad and not too well-defined interdisciplinary subject include: "Biophysical Science," by E. Ackerman, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962; "Biophysics—Concepts and Mechanisms," by E. J. Casey, Reinhold Publishing Corp., New York, 1962; "Molecular Biophysics," by R. B. Setlow and F. C. Pollard, Addison-Wesley Publishing Co., Inc., Reading, Mass., 1962; "Mathematical Biophysics. Physico-mathematical Foundations of Biology," by N. Rashevsky, 3rd Revised Edition, Vol. I, Dover Publications, Inc., New York, 1960.

E. J. CASEY

Cross-references: COLOR, MATHEMATICAL BIOPHYSICS, RADIOACTIVITY, VISION AND THE EYE.

BOLTZMANN'S DISTRIBUTION LAW

Let us consider a system composed of molecules, of one or more kinds, able to exchange energy at collisions but otherwise independent of one another. Evidently we cannot say anything useful or interesting about the state of a particular molecule at a particular time. We can however make useful statements about the average fraction of molecules of a given kind in a given state, or, what is the same thing, the fraction of time spent by each molecule of a given kind in a given state. If the system is maintained at a definite temperature, for example by keeping it in a thermostat, then the fraction of molecules of a given kind in a given state is determined by the energy of this state and by the temperature. In particular if we denote by i and k two completely defined states of a molecule of a given kind and by E_i and E_k the energies of these two states then the average numbers N_i and N_k of molecules in these two states are related by

$$N_i/N_k = \exp \{ -\beta(E_i - E_k) \} \quad (1)$$

where β is a parameter having a positive value determined entirely by the thermostat; i.e., β has the same value for all states of a given kind of molecule and for all kinds of molecules. In other words β has all the characteristics of temperature except that it decreases as temperature increases. If we write

$$\beta = 1/kT \quad (2)$$

then it can be shown that T is identical with thermodynamic (or absolute) temperature and k is a universal constant whose value determines the unit of T called the degree. When k is given the value 1.38041×10^{-23} joules/degree, the temperature scale becomes the Kelvin scale defined by $T = 273.16^\circ\text{K}$ at the triple point of water. Substitution of Eq. (2) into Eq. (1) leads to

$$N_i/N_k = \exp \{ -(E_i - E_k)/kT \} \quad (3)$$

This fundamental relation is called Boltzmann's distribution law after the creator of STATISTICAL MECHANICS, Ludwig Boltzmann, (1844/1906) Professor of Physics in Leipzig, and k is called Boltzmann's constant (see also KINETIC THEORY).

We must now discuss the meaning of the words used above, "completely defined state." These words have one meaning in classical mechanics and a different, but related, meaning in quantum mechanics. Since the quantal definition is the simpler we shall discuss it first. We begin by considering a system of highly abstract "molecules" having only a single degree of freedom, for example linear oscillators. The quantum states form a simple series specified by consecutive integers called the quantum numbers. In this simple example there is no ambiguity in the meaning of "completely defined state"; each state i is completely defined by the integral value of a single quantum number. Let us now consider a "molecule" with three degrees of freedom such as a structureless particle moving in three-dimensional space. The complete specification of this particle's state requires not one but three integral quantum numbers. If the particle moves freely in a cubical box, the three quantum numbers may be associated with motion along the three directions normal to the faces of the box. The subscript labels i and k in the previous formulas are abbreviations for sets of three quantum numbers. For example i might mean (2, 5, 1) and k might mean (3, 4, 2). There can now be several states having the same energy. For a particle moving freely in a cubical box, it follows from symmetry that the states (2, 5, 1), (1, 2, 5), (5, 1, 2), (1, 5, 2), (2, 1, 5), and (5, 2, 1) all have the same energy; such an energy level is called sixfold degenerate. (One should *not* speak of a p -fold degenerate state, but of a p -fold degenerate energy level.) It is sometimes desirable to consider the fraction of molecules having a given energy rather than the fraction in a given state. If N_i and N_k denote the average number of molecules of a given kind having energy E_i and E_k then evidently

$$N_i/N_k = (p_i/p_k) \exp \{ -(E_i - E_k)/kT \} \quad (4)$$

Alternatively, if f_i denotes the average fraction of

molecules of a given kind having energy E_r , then

$$f_r = p_r \exp(-E_r/kT) / \sum_s p_s \exp(-E_s/kT) \quad (5)$$

The sum \sum_s occurring in the denominator is called the partition function.

It may happen that certain degrees of freedom are completely independent of other degrees of freedom. We call such degrees of freedom "separable." The partition function can then be separated into factors relating to the several sets of separable degrees of freedom, and Boltzmann's distribution law is applicable separately to each set of separable degrees of freedom. For example, for an electron moving freely in a rectangular box, the translational motions normal to the three pairs of faces and the fourth degree of freedom due to spin are all separable.

We shall now consider briefly the meaning of completely specified state according to classical mechanics. We know that classical mechanics is merely an approximation, sometimes good but sometimes bad, to quantum mechanics. Motion in each separable degree of freedom can be described classically by a coordinate x and its conjugate momentum p_x . If x and p_x are plotted as Cartesian coordinates, the diagram is called the phase plane. There is a simple correlation between the quantal and the classical descriptions: the density of quantum states is one per area h (Planck's constant) in the phase plane. This may be extended to several degrees of freedom. If there are f degrees of freedom, the motion is described by f coordinates q_1, \dots, q_f and the conjugate momenta p_1, \dots, p_f . We can imagine these plotted in a $2f$ dimensional Cartesian space called phase space. There is then one quantum state per $2f$ dimensional volumes h^f of phase space. In the classical as in the quantal description there can be degenerate energy values and there can be separable degrees of freedom. The classical description is a good approximation to the quantal description when the spacing between energy levels is small compared with kT . An example of an effectively classical separable degree of freedom is the motion in a given direction of a free particle. If the linear coordinate is denoted by x and the linear momentum by p_x , then the fraction of molecules at a position between x and $x + dx$ and having a momentum between p_x and $p_x + dp_x$ is

$$\frac{\exp(-p_x^2/2mkT) dx dp_x}{\int dx \int_{-\infty}^{\infty} dp_x \exp(-p_x^2/2mkT)} \quad (6)$$

where m is the mass of the particle so that its (kinetic) energy is $p_x^2/2m$. In the classical treatment, the kinetic and potential factors are separable. Consequently the fraction of molecules, anywhere or everywhere, having momentum between p_x and $p_x + dp_x$ is

$$\exp(-p_x^2/2mkT) dp_x / \int_{-\infty}^{+\infty} \exp(-p_x^2/2mkT) dp_x \quad (7)$$

Equation (7) is called Maxwell's distribution law

after Clerk Maxwell (1831/79), Professor of Physics at Cambridge (England), who obtained it in 1860 before Boltzmann in 1871 obtained his wider distribution law. Maxwell derived his distribution law from the conservation of energy together with the assumption that the motion is separable in three mutually orthogonal directions. The latter assumption was violently attacked by mathematicians, but we now recognize that the assumption is both reasonable and true.

In conclusion we must mention that a necessary condition for the validity of Eq. (3), and consequently of other formulas derived from Eq. (3) is that $N_i \ll 1$ for the state (or states) of lowest energy and *a fortiori* for all other states. When this inequality does not hold, Boltzmann's distribution law must be replaced by a more general and more precise distribution law, either that of Fermi and Dirac or that of Bose and Einstein according to the nature of the molecules.

E. A. GUGGENHEIM

References

- Guggenheim, "Boltzmann's Distribution Law," Amsterdam, North Holland Publishing Company, 1955.
Mayer and Mayer, "Statistical Mechanics," New York, John Wiley & Sons, 1940.
Tolman, "The Principles of Statistical Mechanics," London, Clarendon Press, 1938.

Cross-references: BOSE-EINSTEIN STATISTICS, FERMI-DIRAC STATISTICS, KINETIC THEORY, QUANTUM THEORY, STATISTICAL MECHANICS, RADIATION, THERMAL.

BOND, CHEMICAL

The chemical bond is a concept used to account for the close association of atoms in neutral or charged molecules. In molecules with more than two atoms, the concept attempts to describe the stability and other properties in terms of the aggregate effect of several bonds. The description of the bonds may then be either *localized*, i.e., confined to attraction between pairs of atoms, or *delocalized*, where the attraction is thought to extend over larger groups of atoms. Various idealized types of bonds are customarily recognized: *ionic* or electrostatic bonds, *covalent* or electron-pair bonds, and metallic bonds.

Ionic bonding exists, e.g., in solid and fused salts. Electrons here are localized in one or the other ion giving rise to charged particles whose mutual attractions and repulsions can be calculated from simple electrostatic theory. Since electrostatic forces do not diminish rapidly with distance, significant interactions still exist beyond nearest ionic neighbors in the crystal lattice. Nevertheless, the *lattice energy*, or the total ionic bonding energy referred to an ideal ionic gas as zero, can be calculated as the sum of the Coulomb energy terms for all possible ion pairs at their equilibrium distances. Empirical corrections for short-range repulsions between electronic envelopes of the

ions must be made, and polarization and van der Waals energies can be taken into account.

Covalent bonding is described qualitatively according to the ideas of Lewis and Kossel as being due to the sharing of electron pairs between bonded atoms. The valence-bond treatment of Heitler and London, elaborated by others, applies quantum mechanical theory to this idea. Approximate wave functions for bonding electron pairs are constructed as products of atomic wave functions. The energy of the covalent bond, referred to the isolated atoms as zero, can then be calculated. A major portion of this energy is the quantum mechanical *exchange energy*, which arises from the fact that either electron may reside on either atom with equal probability.

In certain molecules, e.g., benzene, two or more plausible pairing schemes can be combined in the same calculation; bond energies calculated this way indicate molecular stability is enhanced by the so-called *resonance energy*.

Polar covalent bonds, where electron pairs are not shared equally, are described by including ionic wave functions in the calculation. Among chemical substances, one can find a more or less continuous variation of bond polarity, ranging from purely ionic bonds to nonpolar covalent bonds. The polarity of a bond can be judged qualitatively by the electronegativity difference of the bonded atoms.

Alternately, covalent bonds may be described in terms of the molecular orbital (MO) method, in analogy to the description of electrons in atoms by means of atomic orbitals. Electrons are assumed to be delocalized throughout the molecule consistent with the overall molecular geometry, their motion being described by means of wave functions obtained as linear combinations of atomic orbitals (LCAO approximation). Each molecular orbital can accommodate two electrons with opposed spins and corresponds to a certain level of binding energy for the electrons. The most stable electron configuration corresponding to the molecule in the ground state (lowest energy) is then obtained by filling of MO's with electron pairs in the order of increasing energy.

When there are one or more valence electrons per atomic valence orbital, the description can often be reduced to an equivalent series of MO's localized between two atoms, and thus represents conventional covalent bonds. On occasion, some bonding electrons still remain delocalized as in benzene, the boron hydrides, or carbonate ion; in such situations, molecules tend to be more stable than expected from a conventional view as a consequence of the *delocalization energy*. In this context, the delocalization energy corresponds to the resonance energy of the valence bond description. Molecular orbitals may be classified by the symmetry of the electron distribution in them, e.g., as σ bonds (bond direction is an axis of symmetry) or as π bonds (one symmetry plane containing the bond direction). Multiple bonds always consist of one σ bond and one or two π bonds. Another distinction is based on relative energy: a bonding MO has

lower energy than the constituent atomic orbitals, an antibonding MO has higher energy.

In metals where there are commonly few valence electrons per atom, the MO description is completely delocalized. The valence MO's form, on an energy scale, bands of closely spaced energy levels many of which have the same energy (degenerate levels). For lack of electrons, the bands are only partially filled, the electrons are not all paired and are free to move throughout the solid in this so-called conduction band. In semiconductors, the valence band is completely filled but an empty band, slightly higher in energy, can receive electrons by thermal excitation or other energy input, thus becoming a conduction band. Bonding in semiconductors can be viewed as intermediate between the metallic and the covalent type. Graphite, condensed aromatic hydrocarbon derivatives and molecules with conjugated multiple bonds can also be placed in this intermediate category as evidenced by recent research. Examples of bond type intermediate between metallic and ionic are the alloys and interstitial compounds of transition metals and hydrogen, carbon and nitrogen, and perhaps the nonstoichiometric oxides and sulfides of the transition metals.

Properties which have found use in the systematic discussion of bonding are the strength, length, force constant, and polarity of bonds. The strength of chemical bonds is measured in terms of *bond energies* and bond dissociation energies. Bond energies are defined so that the sum of the energies of the bonds broken in completely atomizing a molecule or complex ion is equal to the energy of atomization of the substance. Bond energies depend on the order of the bond and typically range from 30 to about 100 kcal/mole for single bonds, up to about 150 kcal for double bonds and up to about 220 kcal for triple bonds. Bonds of a given type between a fixed atom and a series of other atoms generally decrease in strength as the principal quantum number of the valence electrons of the other atoms increases.

The *bond dissociation energy* is defined as the energy required to break one given bond in a molecule yielding, in general, atoms and/or smaller molecules. Bond dissociation energies and bond energies are identical only in diatomic molecules. Significant differences in more complex cases may occur, particularly when delocalization or resonance is involved in the broken bond, or in the molecular fragments. This implies that the energy of given bond may depend on the environment in the remainder of the molecule. Nevertheless, the concept has proved invaluable in the correlation of molecular structure with thermodynamic and kinetic stability. Bond energies and bond dissociation energies are derived from heats of formation and reaction, activation energies in reaction kinetics, appearance potentials in mass spectrometry, and from electronic spectra.

Bond lengths, or equilibrium internuclear distances, are determined from x-ray, electron, and neutron diffraction patterns, from rotational spectra and the fine structure of Raman and

infrared spectra. They depend on the identity of the bonded atoms and are a function of the bond strength and the bond polarity. Sets of atomic radii have been developed for the various bond types such that radius sums duplicate normal bond lengths. Comparisons of computed and observed bond lengths can thus be made to characterize the bonds. Bond stretching *force constants* are obtained from vibrational spectra and serve as an additional index of bond character. Force constants are a measure of the curvature of the potential energy surface at the equilibrium distance; they increase with bond strength and decrease with bond length.

Bond dipole moments are a quantitative measure of the polarity of bonds and permit calculation of the distribution of charge between bonded atoms, if bond lengths are known. Bond dipole moments are estimated by expressing the molecular electric dipole moment (obtained from dielectric or Stark effect measurements) as the vector sum of the constituent bond moments. This method should be used with caution since nonbonding electron pairs in directional hybrid orbitals may make significant contributions to the molecular moment.

G. E. RYSCHKEWITSCH

References

- Ryschkewitsch, G. E., "Chemical Bonding and the Geometry of Molecules," New York, Reinhold Publishing Corp 1963.
 Gray, H. B., "Electrons and Chemical Bonding," New York, W. A. Benjamin, Inc., 1964.
 Cartwell, E., and Fowles, G. W. A., "Valency and Molecular Structure," New York, Academic Press, 1956.
 Pauling, L., "The Nature of the Chemical Bond," Third edition, Ithaca, N.Y., Cornell University Press, 1960.
 Coulson, C. A., "Valence," London, Oxford University Press, 1952.

Cross-references: DIELECTRIC THEORY, ENERGY LEVELS, MOLECULES AND MOLECULAR STRUCTURE, POLAR MOLECULES.

BOSE-EINSTEIN STATISTICS AND BOSONS

Bose-Einstein statistics is a type of quantum statistics concerned with the distribution of particles of a particular kind among various allowed energy values taking into account the quantization of the energy values. Quantum statistics is a branch of STATISTICAL MECHANICS which treats the average or statistical properties of a system composed of a large number of particles using standard mathematical techniques and the properties of the constituent particles. It is different from classical statistical mechanics only in that the particles of the system are described quantum mechanically.

Let us consider a system of N non-interacting particles. Three different distributions of the

particles among the various energy levels are possible depending upon the assumptions that are made about the particles. If it is assumed that each arrangement or distribution which conserves energy is equally probable and also that the particles are distinguishable, and if each permutation of particles among the possible levels is counted as a different distribution, one obtains an average for the relative number of particles in the various levels known as the Maxwell-Boltzmann distribution. If the particles are treated as indistinguishable and only the number of different combinations of particles is counted, the Bose-Einstein distribution is obtained. A third distribution, known as the Fermi-Dirac distribution, results if, in addition to indistinguishability, it is required that the particles obey the Pauli exclusion principle which permits no more than one electron in each quantum state.

These three distributions may be expressed mathematically as follows where $n(\epsilon)$ gives the number of particles per energy level at energy ϵ when the particles are in thermal equilibrium at temperature T :

$$(1) \quad n(\epsilon) = \frac{1}{e^{(\epsilon - \epsilon_0)/kT}} \quad \text{Maxwell-Boltzmann}$$

$$(2) \quad n(\epsilon) = \frac{1}{e^{(\epsilon - \epsilon_0)/kT} - 1} \quad \text{Bose-Einstein}$$

$$(3) \quad n(\epsilon) = \frac{1}{e^{(\epsilon - \epsilon_0)/kT} + 1} \quad \text{Fermi-Dirac}$$

where k is the Boltzmann constant and ϵ_0 is related to the number of particles present and depends on the temperature in such a way that for energies large compared to kT (so that the probability of occupation for a level becomes considerably less than unity), all three distributions reduce to the Maxwell-Boltzmann distribution.

The appropriate form of statistics to apply to an assembly of particles can also be discussed in terms of the symmetry properties of the wave functions describing the particles. Two classes of wave function ψ (a solution of the SCHRÖDINGER EQUATION for two or more identical particles) result from interchanging all the coordinates, both spatial and spin, in the wave function. It should be noted that this symmetry class does not change as a function of time. The wave functions for particles obeying Fermi-Dirac statistics (fermions) are antisymmetric, while those for bosons (Bose-Einstein statistics) are symmetric. Therefore, for a system of bosons, if all the coordinates of any pair of identical particles are interchanged in the wave function, the new wave function will be identical with the original.

Photons, pi mesons, and all nuclei of even mass number are bosons, while nucleons, electrons, neutrinos, mu mesons and all nuclei of odd mass number are fermions. All known bosons have angular momentum $nh/2\pi$, where n is an integer or zero and h is Planck's constant. The statistics of some nuclei have been determined experimentally by the observation of the relative intensities of

successive lines in the band spectra of homonuclear, diatomic molecules.

One application of Bose-Einstein statistics to a physical situation is the treatment of a "photon gas." It is possible to obtain the Planck distribution law for blackbody radiation by treating the electromagnetic radiation inside an enclosure at constant temperature as a gas of particles of zero rest mass which obey the Bose-Einstein distribution law. This treatment provides an interesting example of the wave-particle duality found in nature, since it is in marked contrast to the original derivation which was based on the wave nature of electromagnetic radiation.

Another interesting application is the explanation of the superfluid properties of liquid helium which occur below the so-called λ point of 2.18°K. Natural helium is composed mostly of He^4 which has zero spin angular momentum and hence is a boson. A qualitative explanation of the superfluid properties of liquid helium is obtained if care is taken not to follow the usual procedure of assuming that the energy states for the bosons are continuously distributed. At low temperatures, the discrete nature of the lowest levels can be important. If one takes account of the fact that an appreciable number of helium atoms will be in the lowest energy state as the temperature is reduced below the λ point, and if one associates these atoms, which have no thermal energy, with the superfluid component, many of the interesting superfluid properties can be understood. It is significant to note that no such properties are observed for He^3 which is a fermion.

ROBERT L. STEARNS

Cross-references: FERMI-DIRAC STATISTICS AND FERMIONS, PHOTON, SCHRÖDINGER EQUATION, STATISTICAL MECHANICS, SUPERFLUIDITY.

BREMSSTRAHLUNG*

An electron can suffer a very large acceleration in passing through the Coulomb field of a nucleus, and in this interaction the radiant energy (photons) lost by the electron is called bremsstrahlung.¹ (bremsstrahlung† sometimes designates the interaction itself). If an electron whose total energy $E_0 \gtrsim 800/Z$ MeV traverses matter of atomic number Z , the electron loses energy chiefly by bremsstrahlung. This case is considered here.

Bremsstrahlung in the coulomb fields of the atomic electrons is adequately included by replacing Z^2 in the formulas by $Z(Z+1)$. For $Z \lesssim 5$, more complicated correction is required.¹

Protons and heavier particles radiate relatively little because of their large masses (radiation rate is proportional to the square of the acceleration,

inversely proportional to the square of the mass). If a very energetic electron traverses one radiation length (λ_0) of any matter, bremsstrahlung reduces the electron's energy to $1/e$ of its incident value on the average. Some examples are:

Element	Air	C	Al	Fe	Cu	W	Pb
Radiation length λ_0 cm	29800	20	9.1	1.7	1.42	0.32	0.51

The energy dependence of radiation loss by an electron of energy E_0 , traversing matter of density n atoms/cc is given by $dE/dx = -nE_0\phi_{\text{rad}}$ where ϕ_{rad} is given by the curves in Fig. 1

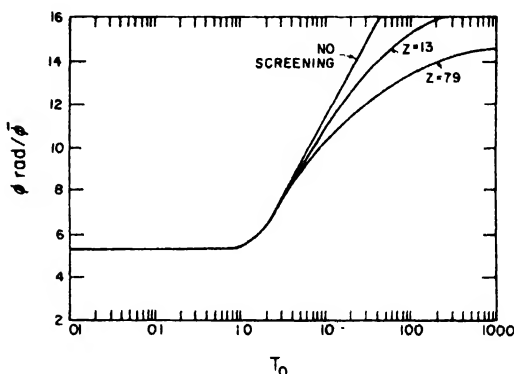


FIG. 1. Dependence of the total radiation cross section $\phi_{\text{rad}} \rightarrow (1/E_0) \int_0^{T_0} k d\sigma_k$ on the initial electron kinetic energy, T_0 . The parameter $\phi = Z^2 r_0^2 / 137$.¹

A beam of energetic electrons incident upon a radiator produces a bremsstrahlung beam that is directed sharply forward. Angular distributions for typical "thick" tungsten targets are shown in Fig. 2. Curves for other heavy elements are similar if all radiator thicknesses are measured in units of the radiation length. In such thick radiators, the incident electrons scatter before radiating appreciably, making any observed distribution actually an average over electron scattering angles of the corresponding basic distribution. The basic angular distribution has a zero at $\alpha = 0$, which is quite different from the curves of Fig. 2. The basic spectral shape is a weak function of photon angle and, in thick radiators, electron scattering modifies this shape slightly (Fig. 3). Examples of thick radiator spectra (bremsstrahlung cross sections) are shown in Fig. 4 for various incident electron energies. The bremsstrahlung spectra depend upon screening of the nuclear coulomb field by atomic electrons through the parameter $\gamma = 51k/[E_0(E_0 - k)Z^{1/3}]$, where k is the photon energy in million electron volts. For complete

* This work was supported by the U.S. Atomic Energy Commission.

† "Bremsstrahlung"—German; bremsen, to brake and Strahlung, radiation.

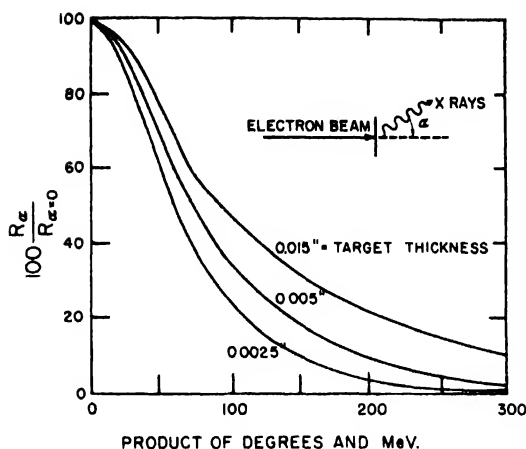


FIG. 2. Theoretical bremsstrahlung angular distributions from thick tungsten targets for relativistic energies. These data are obtained from the Natl. Bur. Std. *Handbook*, 55. R_α is defined as the fraction of the total incident electron kinetic energy that is radiated per steradian at the angle α .¹

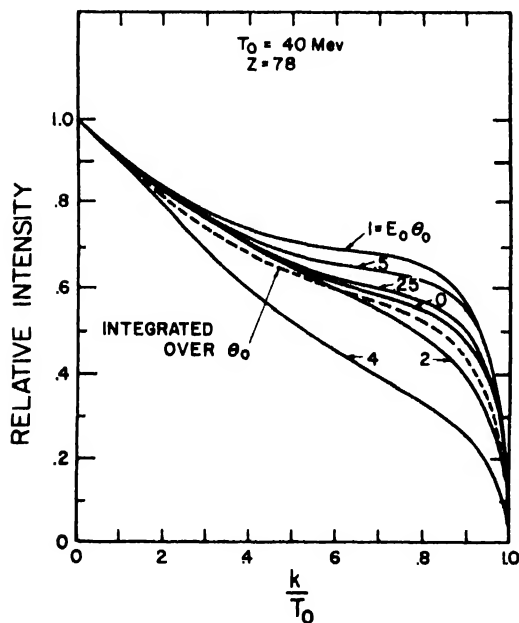


FIG. 3. Dependence of the spectral shape (Schiff's calculation) on the photon emission angle, θ . k is the photon energy in MeV. These curves are from reference 1. T_0 is the incident electron energy in MeV.

screening ($\gamma \approx 0$), the thick radiator spectrum is given by

$$\frac{d\sigma_b}{dk} = \frac{4Z^2r_0^2}{137k} \left\{ \left[1 + \left(\frac{E}{E_0} \right)^2 - \frac{2}{3} \frac{E}{E_0} \right] \cdot \ln(183Z^{-1/3}) + \frac{1}{9} \frac{E}{E_0} \right\} \text{cm}^2/\text{MeV}$$

where E is the final electron *total* energy in million electron volts and r_0 is 2.82×10^{-13} cm. For no screening ($\gamma \gg 1$),

$$\frac{d\sigma_b}{dk} = \frac{4Z^2r_0^2}{137k} \left[1 + \left(\frac{E}{E_0} \right)^2 - \frac{2}{3} \frac{E}{E_0} \right] \cdot \left[\ln \frac{2E_0E}{0.51k} - \frac{1}{2} \right] \text{cm}^2/\text{MeV}$$

Intermediate screening ($2 < \kappa < 15$) leads to much more complicated formulas.¹

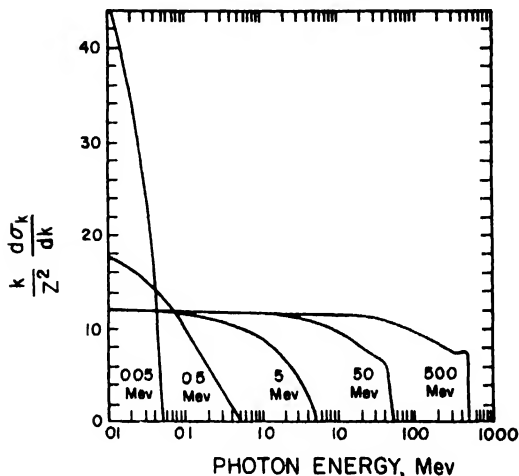


FIG. 4. Dependence of the Born-approximation absolute cross section integrated over the photon directions on the photon and electron energy. These curves are from reference 1.

A remark concerning formulas is in order. Generally, expressions for a given cross section are very different depending upon whether the electron energy is small or very large, upon whether the screening is zero, intermediate, or complete, and upon whether one is dealing with the most usual electron-nucleus collisions or with purely electron-electron collisions. Most calculations have been done in Born approximation. The reader is referred to Koch and Motz¹ for an excellent review article on the subject.

The absolute number of bremsstrahlung photons in the photon energy interval dk radiated by a single electron of energy E_0 traversing a radiator of thickness dt and n atoms/cm³ is given by $(d\sigma_b/dk)n dt dk$, where $d\sigma_b/dk$ can be found from Fig. 4.

It must be noted that photon-electron showers begin developing in approximately one radiation length, and these formulas and curves apply only to the basic bremsstrahlung interaction itself or to radiators somewhat thinner than one radiation length.

Bremsstrahlung beams are partially polarized only from extremely thin radiators ($< 10^{-3}$ radiation lengths) because the angular region of

polarization is sharply peaked about the angle $\theta = m_0 c^2 / E_0$. Electron scattering in the radiator broadens the peak and shifts the maximum to larger angles. Polarization is defined by

$$P(\theta, E_0, k) = \frac{d\sigma_+(\theta, E_0, k)}{d\sigma_-(\theta, E_0, k) + d\sigma(\theta, E_0, k)}$$

where an electron of energy E_0 radiates a photon of energy k at angle θ , and \pm directions are with respect to the plane defined by the incident electron and the radiated photon. When the electron is relativistic before and after the radiation, the electric vector is most probably in the \perp direction. Polarization is difficult to observe because thin, low-yield radiators are required. Practical thick-target bremsstrahlung shows no polarization effects whatever. One usually deals with this unpolarized bremsstrahlung and therefore averages over all possible states of polarization of the incident photons. It is clear that if one makes use of polarized photons when investigating electromagnetic interactions, additional information can be obtained.

ROBERT W. KENNEY

References

1. Koch, H. W., and Motz, J. W., *Rev. Mod. Phys.* **31** 920, 1959. Extensive survey of formulas and excellent presentation of curves for numerical calculation.

Cross-references: ATOMIC AND MOLECULAR BEAMS, COLLISIONS OF PARTICLES

BROWNIAN MOTION

Brownian motion is the randomly agitated behavior of colloidal particles suspended in a fluid. The phenomenon is named for its discoverer, Robert Brown, an English botanist. In 1828 he observed the "perpetual dance" of microscopic pollen grains suspended in water. Initially, this effect was interpreted as being due to the motions of living matter, but it was later found that *any* tiny particles in suspension exhibit Brownian motion.

In 1888, M. Gouy attributed the motion to the bombardment of the visible particles by invisible thermally excited molecules of the suspension. In 1900, F. M. Exner expressed the view that the kinetic energy of the visible particles must equal that of the surrounding suspension particles, and he attempted to estimate molecular velocities on this basis.

In a series of papers published from 1905 to 1908, Einstein¹ successfully incorporated the

suspended particles into the molecular-kinetic theory of heat. He treated the suspended particles as being in every way identical to the suspending molecules except for the vast difference of their size. He set forth several relationships which were capable of experimental verification and he invited experimentalists to "solve" the problem.

Several workers undertook this task. The most notable of these was Perrin.² Perrin's special success was due to his technique of preparing particles to suspend which were of uniform and known size. The uniformity was achieved by fractional centrifuging, and the size was established by noting that they could be coagulated into "chains" whose length could be measured and whose "links" could be counted. The microscopic observation of these uniform particles enabled Perrin and his students to verify the Einstein results and to make four independent measurements of Avogadro's number. These results not only established our understanding of Brownian motion, but they also silenced the last critics of the atomic view of matter.

Probably the simplest example of Perrin's experiments was his test of the Law of Atmospheres. If we assume that the air is at rest and has the same temperature from ground level upward, it can be shown that the pressure (and concentration) of the air falls off exponentially with increasing altitude. For particles of mass m and density ρ suspended in a medium of density ρ' at absolute temperature T , the ratio of the particle concentrations n_1 to n_2 at heights h_1 and h_2 is given by

$$\frac{n_1}{n_2} = \exp \left[- \frac{mg(\rho - \rho')N_0(h_1 - h_2)}{\rho RT} \right]$$

where N_0 is Avogadro's number, g is the acceleration of gravity, and R is the universal gas constant. Although the concentration of air varies slowly with height, the concentration of the relatively heavy particles varied significantly over a height change of a few millimeters. By observing the concentration variation as a function of height, all quantities in the given equation were known except Avogadro's number which could therefore be determined.

JAMES A. RICHARDS, JR.

References

1. Einstein, Albert, "Investigation of the Theory of the Brownian Movement," A. D. Cowper, translator, New York, Dover Publications, 1956.
2. Perrin, Jean, "Atoms," D. L. Hamrick, translator, London, Constable, 1923.

Cross-references: ATOMIC PHYSICS.

C

CALORIMETRY

Calorimetry is the science of measuring the quantity of heat absorbed or evolved by matter when it undergoes a change in its chemical or physical state. The apparatus in which the measurement is performed is a calorimeter, and the experimenter is frequently referred to as a calorimetrist.

When matter is involved in a chemical or physical process, its total energy content is usually altered. The difference in energy between its initial and final states, ΔE , must be transferred to, or from, the environment of the system. This energy exchange between the system and its environment is in the form of heat or work or both. In calorimetry, the energy exchanged as heat is quantitatively evaluated. The heat absorbed by the system, q , is related to the work done by the system on its environment, ω , and the increase in internal (total) energy of the system, ΔE , by the thermodynamic relationship

$$q = \Delta E + \omega \quad (1)$$

When calorimetric measurements are performed at constant pressure and only pressure-volume work is involved, q is equal to the increase in heat content or enthalpy, ΔH . Most calorimetric measurements are performed under these conditions, but when other conditions are imposed, appropriate consideration must be made in the thermodynamic treatment of the data.

The process selected for calorimetric study may be a simple change in the physical state of matter, such as a change in temperature of the material, or it may consist of a series of complex chemical reactions such as are encountered in the combustion of many fuels. In fact, nearly any process involving a chemical or physical change in matter might well become a necessary subject for calorimetric investigation.

Calorimetric determinations of energy changes are essential in many theoretical and practical problems. Heat capacity or specific heat data are vital to the design of heat exchange equipment. The thermal properties of steam and certain metals are a major consideration in the design of modern boilers and turbines. The heats of combustion of fuels are essential in rocket, engine and gas turbine design. The heat liberated by chemical reactions must be considered in the development of chemical process equipment. Often the required

equilibrium constant of a process is most conveniently obtained by a simple calculation from the free energy change, ΔF . For a great many processes, numerical values of ΔF can be obtained from the change in heat content, ΔH , and the entropies of the participating substances, S , using the thermodynamic relationship

$$\Delta F = \Delta H - T\Delta S \quad (2)$$

where T is the absolute temperature. The entropies of the individual substances can generally be evaluated from heat capacity measurements that extend to very low temperatures.

The design and constructional details of calorimeters vary widely because of the diversified nature of the processes suitable for calorimetric study. However, the basic principles are general, and their consideration constitutes a common requirement in practically all designs. Suitable devices and procedures for three essential measurements are usually required, but one or two can sometimes be omitted by operating under certain restrictions. The measurements are: (1) the temperature of the calorimeter and its contents, (2) the quantity of energy that is added to the calorimeter from an external source, and (3) the quantity of heat that is exchanged between the calorimeter and its environment.

Most calorimetric operations involve a temperature change, since the heat liberated (or absorbed) during the process is stored in the calorimeter and its contents by virtue of their combined heat capacity. Thermocouples, thermopiles and resistance thermometers are commonly used for temperature measurements. The quantity of energy liberated or absorbed in a calorimetric process is evaluated in terms of electrical energy. This is done by three similar methods. (1) In an exothermic process where heat is liberated, the calorimeter is cooled to the original temperature; the temperature rise is then duplicated using an electrical resistance heater. (2) The heat absorbed in an endothermic process is supplied by an electrical heater at such a rate as to keep the temperature constant. (3) In heat-capacity measurements, the electrical energy is supplied directly by a heater. Electrical energy and temperature can be measured very accurately by modern methods, but the problem of heat transfer between the calorimeter and its environment is more difficult. The minimization of, and accurate

correction for, heat exchange is the major problem to be reckoned with in modern calorimetry.

When two adjacent bodies (such as a calorimeter and its environment) are not at exactly the same temperature, heat is transferred from the warmer to the cooler body. In calorimetry the transfer is made by three major processes: (1) gaseous convection, (2) conduction, and (3) radiation. Gaseous convection can be completely avoided by evacuating the space between the calorimeter vessel and its environment. When evacuation is impractical, convection can be minimized by suitable geometrical considerations in the design of the calorimeter. It is very important to avoid or at least minimize convection, since the heat transported is a complex function of the temperature difference and an accurate evaluation is impossible. Conduction by air or other gases is also usually minimized by evacuating as much as possible of the space between the calorimeter proper and its environment. Conduction in solid materials, used for supporting the calorimeter and for electrical leads, is minimized by proper choice of materials and geometrical design. For small temperature differences, radiation is usually not a serious problem at low temperatures but is a major contributor to heat exchange at elevated temperatures. Heat exchange by radiation can be limited to a few per cent of the blackbody (maximum) values by the use of suitable reflecting surfaces on the outside of the calorimeter and on the adjacent environment. In the absence of convection and for small temperature differences, the heat transferred, Q , is essentially proportional to the temperature difference, ΔT , and time, t , in accordance with Newton's law of cooling.

$$Q = k\Delta Tt \quad (3)$$

It is apparent that anything that can be done in calorimeter design and operation to minimize the terms on the right hand side of Eq. (3) will aid in decreasing the quantity of heat exchanged. Adiabatic calorimeters are operated on the principle that there is no heat exchange and thus no correction to evaluate, if the calorimeter and its environment are maintained at the same temperature.

In calorimeters containing liquids, there is a possibility of a fourth mechanism for transporting heat. This method involves the transport of matter from the calorimeter and its subsequent condensation on the surrounding surfaces. The effect can be avoided by keeping the environment warmer than the liquid or by completely enclosing the liquid. However, even in a completely enclosed system the possibility of vaporization into the space above the liquid with increasing temperature must be considered for volatile liquids.

There are many different varieties of calorimeters, each being particularly suited for a specific type of measurement. Some general features of several representative types are discussed below.

Low-temperature calorimetry has become an important source of heat capacity data for the

evaluation of entropies of substances from measurements extending from near the absolute zero to room temperature or slightly above. The calorimetric vessel consists of a vacuum-tight metal container in good thermal contact with an electrical resistance heater and a thermocouple or resistance thermometer. The sample under study is sealed in the container along with a small amount of gaseous helium. The helium aids in attaining thermal equilibrium at low temperatures because of its high thermal conductivity. The calorimetric vessel is suspended in an evacuated chamber by some material, such as a strong thread, having low thermal conductivity. This chamber is often within a massive copper block which provides a uniform and stable thermal environment. The temperature of the protective block is kept at a temperature near that of the calorimetric vessel. The heat exchanged is evaluated by observing the temperature difference, ΔT , as a function of time and applying Eq. (3) in an integrated form. The constant, k , is evaluated by observing the change in temperature of the calorimeter vessel and its contents under equilibrium conditions. During this rating period, the temperature change is due entirely to heat exchanged with the environment. Some calorimetrists use the adiabatic principle and maintain the temperature of a protective shield as near as possible to that of the calorimeter. This procedure results in the elimination of heat exchange corrections but is not entirely free from objections. Although low-temperature calorimeters are used chiefly for heat capacity determinations, heats of transition, heats of fusion, and heats of vaporization are also measured.

The dropping method is the most common of the accurate high-temperature procedures for measuring heat contents. This apparatus consists of a carefully regulated furnace and a suitable calorimeter, such as a Bunsen ice calorimeter, operating near room temperature. The sample under investigation is sealed inside of a container that will not undergo chemical reaction at the highest temperature of the measurements. The sample and container are thermally equilibrated with the furnace and then dropped into the calorimeter. The empty container is studied in an identical manner and the difference in the two measurements gives the heat content of the sample relative to the room temperature reference. Heat capacities are derived from a series of such measurements as a function of temperature and the thermodynamic relationship:

$$C_p = \frac{\partial(H)}{(\partial T)_p} = \left[\frac{(\partial H}{\partial T} - H_0) \right]_p \quad (4)$$

where C_p is the heat capacity at constant pressure, H the heat content, H_0 the heat content at the reference temperature, and T the absolute temperature.

The Bunsen ice calorimeter is an example of an isothermal calorimeter that is operated at a fixed temperature. The calorimeter is usually surrounded by ice, making it also adiabatic and thus

free from heat exchange. Bunsen's design makes use of the very large difference between the specific volume of ice and water. The calorimeter contains a closed chamber which is full of ice and water. A pool of mercury is maintained in the bottom of the chamber, and as the ice melts, additional mercury enters and keeps the chamber full. The calorimeter has a universal calibration in the form of energy per unit mass of mercury. In early versions, the quantity of ice melted was used as a measure of the heat liberated in the calorimeter. By replacing the ice with other suitable substances, the restriction of operating at one fixed temperature can be removed.

Quantitative measurements of the heat liberated (or absorbed) during the solution of a solid or of another liquid by a solvent are performed in solution calorimeters. Heats of solution, dilution and mixing are common determinations of this type. In addition to participating in the process under investigation, the solvent is used as a means of attaining uniform temperature and composition throughout the calorimeter. This feature necessitates stirring, which is usually accomplished with mechanically or magnetically driven stirrers. Sometimes, however, the calorimeter itself is rotated. Regardless of the method used, the quantity of heat introduced by the stirring must be determined either directly or indirectly and a suitable correction must be applied. Another feature characteristic of solution calorimeters is the method of adding the sample. Either it must be equilibrated with the solvent in the calorimeter, or its heat content relative to the calorimeter temperature must be determined. A common method for solids is immersing a capsule containing the sample in the solvent and breaking it at the desired time.

The heat of combustion of fuels and similar materials is usually measured by bomb calorimetry. The solid or liquid sample is contained in a bomb (pressure vessel) containing excess oxygen or other suitable gas under pressure. The bomb is immersed in a calorimeter containing a liquid, usually water. The reaction is initiated by igniting the sample with a measured amount of electrical energy, and the heat evolved is measured in terms of the temperature rise of the calorimeter. Electrical energy is usually used to duplicate the temperature rise and thus evaluate the heat liberated. However, sometimes a standard sample of a substance having a known heat of combustion such as benzoic acid, is used to calibrate the apparatus. In bomb calorimetry, corrections to standard conditions must be applied (Washburn corrections) since the system is under pressure and because solutions are usually formed.

There are many other important types of calorimeters, such as flow calorimeters, microcalorimeters, flame calorimeters, etc. Nearly any process can be studied by the investigator who is ingenious enough to devise the appropriate apparatus and who has the resources and patience to undertake an extensive project. Although calorimetric measurements are in general time-consuming and tedious, they are essential for a

fundamental and practical understanding of many important chemical and physical processes.

J. E. KUNZLER

Cross-references: ENTROPY, HEAT CAPACITY, HEAT TRANSFER, THERMODYNAMICS.

CAPACITANCE

Definition and Fundamental (Quasi) Static Properties. If a constant voltage V [V = volts] is applied between two conductors insulated from each other, electrical charges Q [Q = coulomb] are so distributed that the conductors form equipotentials. The measure for the charges stored is the capacitance C [F = farad = $10^6 \mu F$ = $10^9 m\mu F$ = $10^{12} pF$] of the capacitor so formed.

$$Q[As] = C[F] \cdot V[V] \quad (1)$$

It is often more convenient to express this storing capacity in terms of energy

$$E[Ws] = (1/2)V^2[C[F]] \quad (2)$$

C is defined by

$$C = \frac{1}{V} \int i dt \text{ or } C = I / \frac{dv}{dt} \quad (3)$$

For capacitor discharge (E_0 = starting voltage),

$$e_c/E_0 = e^{-t/\tau} \quad (4a)$$

and for capacitor charge (E_0 = battery voltage)

$$e_c/E_0 = 1 - e^{-t/\tau} \quad (4b)$$

with the time constant τ

$$\tau = CR \quad (5)$$

where R is the resistor through which the capacitor is being (dis)charged.

For sinusoidal excitation of angular frequency ω , the reactance of the lossless capacitor is

$$V[V]/I[A] = (-j)X[\Omega] = 1/j\omega C[\Omega] \quad (6)$$

If, in electrical circuits, capacitors are connected in parallel, their capacitances add

$$C = \sum_{k=1}^n C_k \quad (7a)$$

If capacitors are connected in series, their elastances (the reciprocal of capacitance, S) add

$$S = \sum_{k=1}^n S_k \quad (7b)$$

Losses in the dielectric may be expressed by a complex relative dielectric constant

$$\epsilon = \epsilon' - j\epsilon'' \quad (8)$$

where $\epsilon''/\epsilon' = Q_e$ determines the dielectric quality factor. For $Q_e > 10$, the loss resistance of the capacitor is given by

$$(1/\omega C)/r_n = Q_e = R_p/(1/\omega C) \quad (9)$$

where r_s is the equivalent series and R_p is the corresponding parallel loss resistance. $Q_e = 1/DF$ (DF = dissipation factor). The power factor is related to DF by

$$PF = DF \sqrt{1/(1 + DF^2)} \quad (10)$$

The loss factor $\tan \delta$ (DF) $\cdot \epsilon$ is proportional to the energy loss/cycle voltage²/volume

Capacitors are used for: (1) frequency determining or selective networks [LC circuits and filters; cf. Eq. (6)]; (2) energy storage [Eq. (2)], for instance, the capacitor being slowly charged and quickly discharged [Eqs. (9) and (10)] in a short burst of energy; and (3) integrators and differentiators [in conjunction with R ; cf. Eq. (3)].

Geometry. Uniform Fields. For a uniform field as, for instance, given between two closely spaced parallel metallic plates (area A in square meters, distance l in meters) and disregarding edge effects

$$C[F] = \epsilon_0 \epsilon A/l \quad (11)$$

with ϵ_0 dielectric constant of free space = $(36\pi \times 10^9)^{-1}$ [F/m] and ϵ = the relative dielectric constant (dimensionless) of the material between the plates.

Discontinuity in Uniform Fields. If, in the above case, the dielectric consists of two sheets of different materials with ϵ_1 (having thickness l_1) and ϵ_2 (having thickness l_2)

$$\frac{E_1}{E_2} = \frac{\epsilon_2}{\epsilon_1} \quad (12)$$

where E_n is electric field strength V_n/l_n . (13)

Equation (12) is of great practical significance if one of the ϵ 's is very high, since then the sheet with the low ϵ carries nearly all voltage (for this reason, for example, higher- ϵ ceramic capacitors have to have fired-on electrodes).

Nonuniform Fields. The most common capacitance with nonuniform fields is the coaxial capacitor (inside diameter d , outside diameter D). Its capacitance is

$$C'[\text{pF/m}] = 55.6 \epsilon / \ln(D/d) \quad (14)$$

Extreme cases of nonuniformity, often causing corona, exist on the sharp edges of plate capacitors. Remedy: For field equalization, deform plates to follow equipotential lines of half potential in a capacitive field with twice the spacing of the original, flat plates (Rogowski profile).

Dielectrics. The dielectric "constant" is often not constant but a function of crystal orientation (anisotropy), temperature, voltage, and frequency (dispersion).

The objective of developing a good fixed capacitor is to have the largest capacity in the smallest possible volume for a given operating voltage. Ideally, the capacitance is not to change with voltage, temperature, time, mechanical stress, humidity, and frequency, and (in most cases) is to have a minimum of losses. The greatest capacitance can be achieved by maximizing ϵ (Case a) and A (Case b), and minimizing l (Case c) [cf. Eq. (11)].

Typical for Case (a) are ceramic capacitors made in discoidal or tubular form (and now recently also as coaxially laminated capacitors). There are four classes of ceramic dielectrics:

(1) Semiconducting, so-called layered, ceramics with dielectric constants above 10^3 . These can be used only for very low voltages (transistor circuits), are quite lossy, and have a strong dispersion of ϵ in the megacycle range.

(2) High- ϵ' dielectrics (mostly barium titanates) with ϵ' in the order of 6000. These are quite temperature- and voltage-sensitive (nonlinearity and hysteresis) and are used as guaranteed-minimum-value capacitors (GMV).

(3) So-called stable dielectric capacitors with an ϵ' of 2000 or, if doped with rare-earth materials, with an ϵ' of 3000 to 4000. These are much less dependent on temperature and applied dc voltage.

(4) Linear, high- Q (in the order of several thousand) temperature-compensating capacitors made with a prescribed (P positive, N negative, or NPO) temperature coefficient of the capacity for incorporation in temperature-stable tuned circuits (compensation of the temperature coefficient of the inductance). The ϵ of such materials lies between 10 and 100.

Case (b) (large A) is exemplified best by stacked plates [silvered mica (for military use; excellent Q , temperature coefficient about ± 100 ppm) or ceramic (monolithic)] or rolled dielectric strips [polystyrene (excellent Q ; commercial use; also about ± 100 ppm T.C.); "Mylar"; oil-impregnated paper; "Teflon" etc.].

Case (c) (small l) is represented by polarized capacitors (to make them unpolarized, two capacitors are connected in series in polarity opposition, usually in the same housing), and it includes the older, larger, and cheaper types like the aluminum foil electrolytics. The newer, more costly, but much smaller, types (having much less leakage current) are tantalum oxide capacitors. Ta_2O_5 stands continuously the extraordinary field strength of $3 \cdot 10^6$ V/cm with an ϵ' of 25, l being measured in angstroms. The Q is about 100. For microminiaturization, silicon monoxide or dioxide or tantalum oxide films of very small l are utilized.

Rating. The reliability of a capacitor is predominantly determined by the dielectric and the seal of the housing. One has to distinguish between failure value and withstand value. The failure value of dielectric strength is the voltage at which the material fails and is conventionally given as the average failure voltage. In contrast, the withstand value is a voltage below which no failure can be expected.

Deterioration of capacitors with time (aging) can be greatly reduced by systematic "physics of failure" investigations. Typical failure mechanisms are, for instance, precorona discharge in adsorbed air layers, or silver migration.

Distributed Capacitance. At higher frequencies, the static and quasi-stationary relations listed on p. 89 become useless, and a more rigorous treatment using Maxwell's equations is necessary.

At very high frequencies, capacitors no longer behave like capacitors unless properly modified (see reference 1). Stray capacitances between windings of coils may make the coil resonate and act capacitively beyond the resonance frequency.

Nonlinear Capacitance. The dielectric of highly nonlinear capacitors is the depletion layer formed at the p - n junction by application of proper bias. These back-biased diodes have a reasonable Q and are widely used as nonlinear reactances in parametric amplifiers for VHF and higher frequencies. Nonlinear ceramics are less suitable for this purpose because of their high losses and great temperature dependency.

H. M. SCHLICHE

Reference

- 1 Schlicke, H. M., "Essentials of Dielectromagnetic Engineering," New York, John Wiley & Sons, Inc., 1961.

Cross-references: DIELECTRIC THEORY, POTENTIAL.

CARNOT CYCLES AND CARNOT ENGINES

Description. The Carnot cycle is often represented, as shown in Fig. 1, by a device which uses an ideal gas as its working substance. A quantity

of gas is confined in a cylinder with a wall so well insulated that no heat can flow through it. The cylinder's heat-conducting base rests on the first reservoir, whose constant temperature is T_h , and the gas assumes this temperature. A weighted insulating piston holds the pressure of the gas at P_1 at which pressure its volume is V_1 . The gas is then said to be in thermodynamic state 1 characterized by P_1 , V_1 , and T_h . Little by little the weight on the piston is now set aside until the pressure is reduced to P_2 and the volume is expanded to V_2 . The gas is now in state 2 characterized by P_2 , V_2 , and T_h . The transition from one thermodynamic state to another is called a thermodynamic process. This is a reversible process if done so slowly that no temperature differences arise within the gas and if the piston moves without friction. The cylinder base was kept at temperature T_h , so during this process the temperature of the gas remained at T_h ; that is, it was an isothermal process. When this is shown on a pressure-volume diagram, the process appears as a portion of the T_h isotherm. To hold the temperature constant, some heat energy Q_h must flow from the reservoir into the gas. In expanding against the weight on the piston, the gas did work W_{12} which is represented on the P - V diagram by the crosshatched area. The cylinder is next moved to an insulated pad where the pressure is further decreased by setting aside more weights,

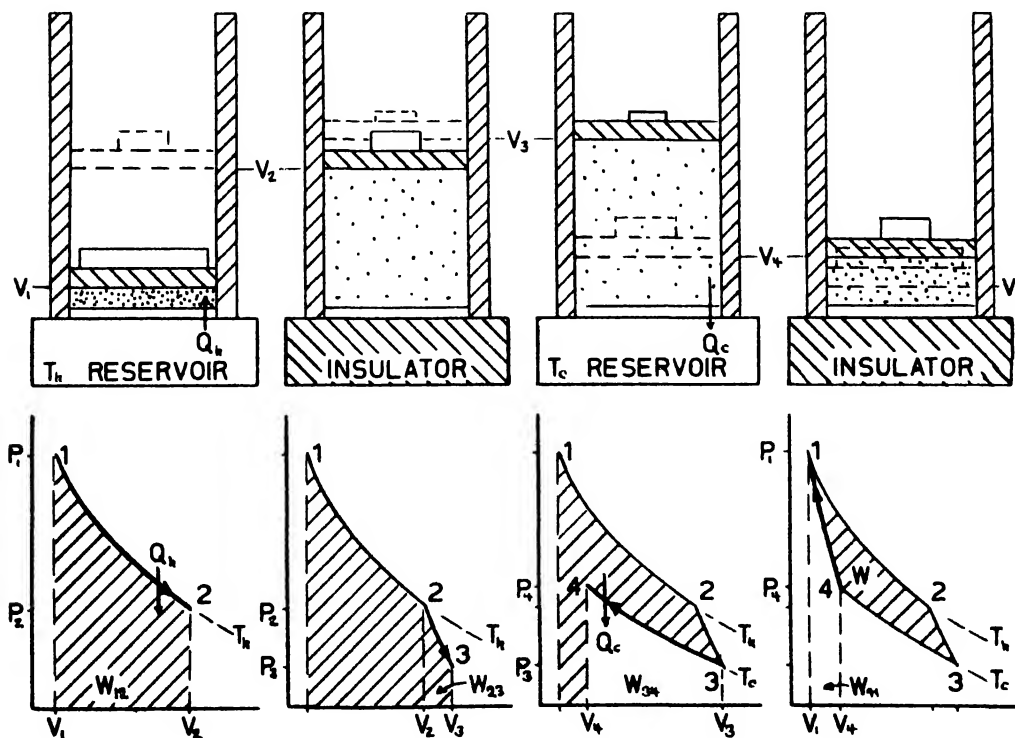


FIG. 1. The steps of a Carnot cycle and the corresponding pressure-volume diagrams. The piston location at the start of each process is shown in solid lines; at the end, in dashed lines. The pressure on the gas can be estimated by the area of the weight shown on top of the piston. The dots indicate that the molecules of the ideal gas are close together when the volume is small and are farther apart when the volume is large.

and the gas again expands. No heat energy flows into the gas from the outside during this expansion, and such a process is called an adiabatic process. The temperature decreases and when it reaches T_c , the temperature of the second reservoir, the process is stopped. The gas is now in state 3 characterized by P_3 , V_3 , and T_c . During this process the gas has done work W_{23} against the load. The cylinder is then moved to the second reservoir where enough weights are slowly replaced to bring the gas to state 4, the point on the T_c isotherm from which state 1 can be reached by an adiabatic process. During this isothermal compression, heat energy Q_c flows from the gas into the reservoir, and the piston does work W_{34} on the gas. To show that this work is done on the gas while previously the work was done on the piston, the appropriate portion of the previous crosshatched area has been removed. The cylinder is finally placed on an insulated pad, the remaining weights are slowly added, and the gas returns adiabatically to state 1. Again the fact that the piston does work W_{41} on the gas, is shown by the removal of the cross-hatched area. One Carnot cycle is now completed.

Definition and Characteristics. A Carnot cycle is any reversible cyclic thermodynamic operation composed of four processes which are alternately isothermal and adiabatic. Since no natural process is strictly reversible, the Carnot cycle is an idealization.

The gas is unchanged by the cycle, but heat energy has flowed and work has been done. Heat energy Q_h was removed from the hotter reservoir, and a smaller amount Q_c flowed into the cooler reservoir so that the heat energy budget of the gas increased by $Q_c - Q_h = -Q_{net}$. The net work done by the gas on the piston was $W = (W_{12} + W_{23}) - (W_{34} + W_{41})$. The mechanical energy budget of the gas decreased by W which is represented by the area inside the cycle. The first law of thermodynamics requires that $W - Q = 0$ so there is no energy residue left in the gas. If the cycle is traversed as described, heat energy Q_h is removed from the higher-temperature reservoir. Part of this remains in the form of heat energy Q_c as it flows into the cooler reservoir, and part of it is converted to mechanical energy as the work W done on the piston. A Carnot cycle operated in this direction is called a Carnot engine. If the direction of operation were reversed, the cycle would be called a Carnot refrigerator. In such a device mechanical energy, the work W done by the piston, is converted into heat energy which combines with the heat energy Q_c which flows from the cooler reservoir into the gas. All this heat energy Q_h flows out of the gas into the hotter reservoir.

The efficiency η of any engine is the fraction of the heat energy input Q_h which is converted into mechanical energy W ; that is, $\eta = W/Q_h$. The following properties of Carnot engines are derived in many textbooks of thermodynamics.

- (1) The efficiency of a Carnot cycle depends only on the temperatures of the two reservoirs.
- (2) No heat engine operating in cycles between

two reservoirs can have a greater efficiency than a Carnot engine operating between those reservoirs.

The Carnot engine is thus a standard against which other heat engines can be compared. Historically it was the source of a number of ideas that are now basic to the study of thermodynamics.

History of the Carnot Cycle. In 1824 Sadi Carnot¹ (1796–1832) analyzed a heat engine assuming that heat can perform mechanical work in falling from a higher temperature to a lower just as water can do work in falling from a higher level to a lower, and assuming that since no water is destroyed in the activity, no heat should be lost either. (His work preceded by more than 20 years the theory of Joule and Helmholtz on the mechanical equivalence of heat.) In his study he proposed an ideal heat engine with the following characteristics: it operated in a continuous cycle and it was reversible. He then showed that it is impossible in a cyclic operation to obtain work from a single constant-temperature heat source and that no more work can be obtained from any process than is required to reverse it.

His ideas escaped notice until 1834 when Clapeyron² recognized their merit, suggested some of the details of the device described above, and plotted its behavior on a P - V diagram. Again the ideas were neglected until William Thomson (later Lord Kelvin) learned of Carnot's work through Clapeyron's memoir. In 1848, Thomson described³ how a Carnot engine could be used to define a temperature scale that was absolute in the sense that it did not depend on what thermometric substance was used. It was based on a series of Carnot engines, each of which did the same amount of work. This was the first important idea drawn from a study of the Carnot cycle. In 1850 Clausius⁴, who learned of Carnot's ideas through Thomson and Clapeyron, showed how Carnot's assumption (no loss of heat) could be reconciled with the newer views of Joule and Helmholtz (which now form the basis of the first law of thermodynamics). It was only required that the engine exhaust less heat energy, by the amount of the work done, than it accepted. Thomson independently reached the same conclusion⁵ in 1851.

In 1854 Clausius⁶ in his study of the Carnot cycle identified the physical property he later named "entropy." This was the second important idea drawn from study of the Carnot cycle. In 1877 Boltzmann⁷ took the principle of Clausius that real processes evolve naturally toward states of higher entropy (which is the second law of thermodynamics) as a basic point in his theory and thus had no need to consider the Carnot cycle from which that principle was derived. It is now common practice to discuss thermodynamics axiomatically rather than historically so that the Carnot cycle no longer plays the important role it once did. The most complete discussions are found in the older volumes,⁸ but some recent books do describe the cycle in detail.⁹ An historical account with an elementary presentation of the theory is also available.¹⁰

ROBERT A. LUFBURROW

References

1. Carnot, S., "Reflexions sur la Puissance Motrice du Feu," Paris, Bachelier, 1824. Reprinted (together with references 2 and 4, below) by Dover Publications, New York.
2. Clapeyron, E., reprinted in *Pogg. Ann.*, **59**, 446-451, 566-586 (1843). See reference 1.
3. Thomson, W., "Mathematical and Physical Papers," Vol. I, pp. 100-106, Cambridge, University Press, 1882.
4. Clausius, R., *Pogg. Ann.*, **79**, 368-397, 500-524 (1850). See reference 1.
5. Thomson, W., *Pogg. Ann.*, **79**, 174-316 (1850).
6. Clausius, R., *Pogg. Ann.*, **93**, 481-506 (1854); **125**, 390 (1865).
7. Boltzmann, L., "Lectures on Gas Theory," (translated by S. G. Brush) Berkeley, University of California Press, 1964.
8. Birtwistle, G., "The Principles of Thermodynamics," Cambridge, University Press, 1925.
9. Preston, T., "The Theory of Heat," Third edition, London, The Macmillan Co., 1919.
9. Shortley, G. H., and Williams, D. E., "Principles of College Physics," Englewood Cliffs, N. J., Prentice-Hall, Inc., 1959.
- Zemansky, M. W., "Heat and Thermodynamics," Fourth edition, New York, McGraw-Hill Book Co., Inc., 1957.
10. Sandfort, J. F., "Heat Engines," Garden City, N. Y., Doubleday & Co., Inc., 1962.

Cross-references: HEAT, THERMODYNAMICS.

CATHODE RAYS. *See* ELECTRON.

CAVITATION

Cavities may form, grow, and collapse in a liquid when variational tensile stresses are superimposed on the prevailing ambient pressure. Pure liquids have theoretical tensile strengths which are estimated on various grounds to be of order 300 to 1500 atmospheres (bars), but the observed tensile strengths of real liquids are much lower. It is presumed, therefore, that the observed tensile strength is a measure of the stress required to enlarge the minute cavities, or cavitation nuclei, which already exist in the liquid rather than the stress required to form new interior interfaces.

The transient cavities formed by tensile stress are unstable and would grow indefinitely if the stress were maintained. After the cavitation nuclei have been expanded to many times their original size, however, they may collapse violently if the stress is reduced or removed. The kinetic energy of the liquid that follows each inwardly collapsing interface becomes highly concentrated as the cavity collapses. If such transient cavities contain very little permanent gas, the peak pressures at collapse may reach thousands of bars, the temperature may reach thousands of degrees, and strong SHOCK WAVES may be radiated to a distance of several cavity radii. Similar cavities formed in saturated liquids will usually contain more gas and their collapse will be less violent,

but the peak pressures attained are still sufficient to produce unique mechanical effects such as the corrosion and pitting of metallic surfaces (as in marine propellers and sonar projectors) and the beneficial removal of embedded dirt (as in ultrasonic cleaners). In the latter case, the soil to be removed provides a prolific source of cavitation nuclei at exactly the sites where cavitation is desired.

In hydrodynamic cavitation, the tensile stress is of relatively long duration and plenty of cavitation nuclei are usually available. As a result, cavitation occurs when the total net pressure, or the stagnation pressure, becomes approximately equal to the vapor pressure of the liquid. In acoustic cavitation, the cyclic pressure required to produce cavitation is a function of the frequency, the partial pressure of any dissolved gas, and the population of cavitation nuclei. For frequencies above about 200 kc/sec, the threshold pressure for cavitation increases with the square of the frequency and is almost independent of the degree of gas saturation. For frequencies below 200 kc/sec, the threshold pressure is a function of the partial pressure of the dissolved gas. In saturated liquids, at sound pressures less than a few bars, stable bubbles can grow from cavitation nuclei by the process of rectified diffusion. At higher levels of acoustic excitation, transient cavities can be formed. The threshold sound pressure at which they appear and the violence of their collapse increase as the partial pressure of the dissolved gas is lowered.

The physical nature of the cavitation nucleus, the details of its dynamic growth from sub-microscopic to visible size, and the peak pressures and temperatures achieved at the climax of collapse are current topics of active research interest.

F. V. HUNT

Reference

- Flynn, H. G., "Cavitation," in Mason, W. P., Ed., "Physical Acoustics," Vol. 1B, Ch. 9, New York, Academic Press, (1964).

Cross-references: ACOUSTICS, ULTRASONICS, LIQUID STATE.

CENTRIFUGE

In most cases, the centrifuge is used for producing sedimentation in fluids, i.e., for the concentration and purification of materials.^{1,2} However, it also has been used extensively as an analytical tool for determining particle or molecular weights and sizes^{1,3,4}; as a means of determining the strength of materials, and numerous other research and practical problems.¹ Centrifugal fields of 10^9 times gravity have been employed in some experiments. The effective centrifugal force, F , on a particle of mass m and density ρ in a fluid of density ρ' is given by the relation $F = m(\rho - \rho')\omega^2 r/\rho$. This force is opposed by the frictional force of the fluid on the particle. If the

speed of sedimentation v is not too large, i.e., the Reynolds number, $v\rho d/\eta$ does not exceed the order of unity, where d is the diameter of the particle and η the coefficient of viscosity, and if the wall effects are neglected, the force of friction f is given by Stokes law. In the case of sedimentation in a liquid for a particle of effective radius a

$$4/3\pi a^3(\rho - \rho')\omega^2 r = 6\pi\eta ar \quad (1)$$

Since ω , r , ρ , ρ' and η are measurable, a and hence the mass of the particle m can be determined. If we are concerned with a substance of molecular weight M , partial specific volume \bar{V} , molar frictional constant f , density of solution ρ , and diffusion constant D , the corresponding equation is³

$$M(1 - \bar{V}\rho)\omega^2 r = f \frac{dr}{dt} \quad (2)$$

For a dilute solution $f = RT/D$ where R is the gas constant and T is the absolute temperature. Hence,

$$v = \frac{dr}{dt} = \frac{MD(1 - \rho\bar{V})}{RT} \omega^2 r \quad (3)$$

The velocity of sedimentation in a unit field $s = \frac{dr/dt}{\omega^2 r}$. If M is known s can be calculated since the other factors in the equation can be determined, or if s is measured, M can be found. The quantity s is called the sedimentation constant and is expressed in Svedberg units (10^{-13} sec)⁻¹ named in honor of the great pioneer in the field of molecular weight measurement by ultracentrifugation. It is a very important quantity for characterizing a substance in solution.^{3,4}

From the above equations alone, one might expect a substance would completely settle out of the solution even with very weak centrifugal fields or gravity alone if the field is applied for sufficient time. This, of course, is not the case because of thermal agitation or BROWNIAN MOTION of the molecules or particles, which gives rise to back diffusion. It can be shown⁵ that the average displacement of a particle in time τ due to Brownian motion is $\Delta X = 2D\tau$ and for a spherical particle of radius a the average velocity for a time τ is

$$v_r = \frac{\Delta X}{\tau} = \left(\frac{RT}{N} \frac{1}{3\pi\eta a\tau} \right)^{1/2}$$

where N is the avogadro number. It is clear that when v_r becomes very much larger than the settling velocity v nothing can sediment out of the solution. When this is the case, it can be shown^{3,4} that when equilibrium is reached between sedimentation and back diffusion

$$\log_e \frac{C_1 f_1}{C_2 f_2} = \frac{M(1 - \rho\bar{V})\omega^2(r_2^2 - r_1^2)}{2RT} \quad (4)$$

where C_1 and f_1 are the concentration and activity coefficient, respectively, at the radius r_1 and C_2 and f_2 are the corresponding quantities at the radius r_2 . If the substance is in dilute solution

in a sector-shaped centrifuge cell, the weight-average molecular weight M_w is given by^{3,4}

$$M_w = \frac{2RT}{(1 - \bar{V}\rho)\omega^2(r_h^2 - r_a^2)} \frac{C_h - C_a}{C_0} \quad (5)$$

where C_a is the concentration at the radius of the meniscus r_a , C_h is the concentration at the peripheral radius r_h and C_0 is the average concentration.

In deriving the above equations, the effect of electrical charges has been neglected. Usually, the solutions are kept near the isoelectric point, but in many important cases this is not possible. The effects of charges on the above equations have been investigated and in some cases found to be quite large.^{3,4,6}

For the separation, purification, or concentration of materials in solution or suspension in a liquid, the centrifugal field is made high enough so that the sedimentation velocity v is appreciable. Equations (1), (2) and (3) are used for estimating the sedimentation. However, these equations hold strictly only when the sedimentation is radial and no turbulence, radial flow, or re-mixing occurs. In most centrifuges used commercially (the cream separator, for example), the liquid flows through the machine during the sedimentation. Any radial flow stream is acted upon by coriolis forces which usually produce mixing. Also, the temperature may not be uniform throughout the sedimenting column. This gives rise to convection if the temperature change produces greater densities near the axis. The driving force which generates thermal convection is roughly proportional to the density gradient times the centrifugal field. Since the latter is large in most centrifuges, the temperature gradient must be small. For a detailed discussion of commercial type centrifuges, both flow-through and batch-type and their operation, reference should be made to Keith and Lavanchy² and others.⁷ In addition to the use of the centrifuge in industry, it is widely used in research and testing laboratories for the purification and preparation of many different substances. As should be expected from theory, high speed centrifuges do not deactivate most molecular species. Even molecular species such as are often encountered in biology and medicine which are stable only over a few degrees of temperature and a small range of pH are not appreciably affected by a comparatively large centrifugal field. For this reason, high speed centrifuges are widely used in biochemistry and molecular biophysics. The rotors of these high speed centrifuges usually spin in a good vacuum (below 10^{-5} torr) to avoid heating and thermal gradients in the rotor. Such centrifuges with push-button control are readily obtainable commercially.

Analytical Centrifuges. The centrifuge is employed as an analytical tool in one of two general methods.^{3,4} The first method makes use of Eqs. (1), (2) and (3), while the second is based upon Eqs. (4) and (5). Sometimes combinations of these two methods are used.^{3,4} When a centrifuge is employed in analytical work, it is usually called

an ultracentrifuge.³ The first method has been more widely used, at least, until recently. In this method, comparatively high rotor speeds with the resulting high centrifugal fields are employed in order to produce an easily measurable rate of sedimentation v . The value of r is usually measured by optical means and s is computed. The high centrifugal field quickly produces a small density gradient across the centrifuge cell which stabilizes the sedimenting column. As a result, very high accuracy in rotor temperature and speed control are not mandatory, although desirable. Another important factor is that the time of centrifugation is comparatively short (~hours). Furthermore, if the solution contains a number of molecular species, each species sediments at its characteristic rate and the value of its sedimentation constant s is easily determined. The effect of ionic charge and the pH of the solution on the values of s and the effect of concentration of one species on the other during sedimentation, etc., have been quantitatively investigated, both theoretically and experimentally, by a number of workers.^{3, 17, 18} The rate-of-sedimentation method has the disadvantage of not being an absolute method and requiring a knowledge of the diffusion constant D as well as the shape of the sedimenting particle or molecule. Often these factors introduce large uncertainties in the value of M .

The second method, known as the equilibrium method, is based upon Eqs. (4) and (5). It is a reliable, absolute method since it is based upon equilibrium thermodynamics. Also, it is not necessary to know the value of the diffusion constant or the shape of the molecule to get the value of M . As pointed out above, no actual sedimentation on the walls of the centrifuge cell occurs in the equilibrium method so that the rotor speed and resulting centrifugal field are relatively low. Consequently, if the concentration of the solute in the solvent is low, the density gradient in the cell is small which in turn makes the sedimenting column sensitive to small rotor temperature and rotor speed variations. Recently, considerable effort has gone into designing and adapting ultracentrifuges to equilibrium measurements and the centrifuging time has been reduced from several days to several hours.^{1, 3, 19} An over-all precision in the measurement of the molecular weights of between 0.1 and 1 per cent over the molecular weight range from 100 to 10⁶ has been obtained.⁹ The determination of molecular weight distributions, etc., can be carried out.

Gas Centrifuging. The centrifuge has been used for removing fine particles suspended in gases and for the separation of gaseous mixtures. The fine particles sediment on the inner wall of the centrifuge where they are removed while the centrifuge is spinning or when it is stopped. In the separation of gases and vapors a tubular type centrifuge is usually used in which the gases flow out of the centrifuge in a light and heavy fraction. Such tubular centrifuges have been used for the separation of isotopes.^{7, 10-12}

When a centrifugal field is applied to a gas it

sets up a pressure gradient

$$dp/dr = Mp\omega^2r = \frac{Mp}{RT}\omega^2r$$

where p is the molar density and P the pressure. In a binary mixture of two ideal gases with molecular weights M_1 and M_2 and with mole fraction N of the lighter gas of molecular weight M_1 , it has been shown both theoretically and experimentally that at equilibrium where r_2 is the radius of the inside periphery of the centrifuge tube

$$\left(\frac{N}{1-N}\right)_{r=r_2} = \left(\frac{N}{1-N}\right)_{r=r_0} \exp \frac{(M_2 - M_1)\omega^2r_2^2}{2RT}$$

In order to determine the value of a centrifuge or cascade of centrifuges for separating isotopes or gases, it is customary to calculate the separative work or separative power which is a measure of separation produced by a single centrifuge or a number of centrifuges used in a cascade. Cohen¹³ has shown that the separative power of a single centrifuge is

$$\delta U = \frac{Dp}{RT} \left[\frac{(M_2 - M_1)\omega^2r_2^2}{2RT} \right]^2 \frac{\pi Z f}{2} \text{ moles/sec}$$

where Z is the centrifuge tube length, D the diffusion constant, and f the flow factor which depends upon the flow pattern in the centrifuge and has a maximum value of one. The number of centrifuges required to carry out a given amount of separation in a time t is $\frac{U}{t\delta U}$ where U is the total separative work and is defined as¹⁴

$$U = W(2N_w - 1) \log_e \frac{N_w}{1 - N_w} + P(2N_p - 1)$$

$$\log_e \frac{N_p}{1 - N_p} + F(2N_F - 1) \log_e \frac{N_F}{1 - N_F}$$

where F is the number of moles of feed material of mole fraction N_F and W and N_w and P and N_p are the corresponding values for the waste and the product. It will be observed that the effectiveness of a centrifuge for gaseous or isotope separation increases directly as the fourth power of the peripheral speed, as the length of the centrifuge, and as $(M_2 - M_1)^2$.

J. W. BEAMS

References

1. Beams, J. W., *Science*, **120**, 619 (1954); *Physics Today*, **12**, 20 (1959) (see for other references).
2. Keith, F. W., Jr., and Lavanchy, A. C., in Kirk-Othmen's "Encyclopedia of Chemical Technology," second edition, New York, Interscience Publishers, 1964.
3. Svedberg, T., and Pederson, K. O., "The Ultracentrifuge," Clarendon Press, 1940.
4. Schachman, H. K., "Ultracentrifugation in Biochemistry," New York, Academic Press, 1959.

5. Burton, E. F., "The Physical Properties of Colloidal Solutions," London, Longmans, Green and Co., 1938.
6. MacInnes, D. A., "Principles of Electrochemistry," New York, Reinhold Publishing Corp., 1939.
7. Beams, J. W., *J. Appl. Phys.* **8**, 795 (1937); *Rev. Mod. Phys.*, **10**, 245 (1938); *J. Wash. Acad. Sci.*, **37**, 221 (1947) (see for other references).
8. Williams, J. W., "Ultracentrifugal Analysis in Theory and Experiment," New York, Academic Press, 1963.
9. Beams, J. W., Boyle, R. D., and Hexner, P. E., *J. Polymer Chem.* **57**, 161 (1962).
10. Beams, J. W., Snoddy, L. B., and Kuhlthau, A. R., *Proc. 2nd U.N. Geneva Conf.* **4**, 428 (1958).
11. Beyerle, K., Groth, W. E., Nann, E., and Welge, K. H., *Proc. 2nd U.N. Geneva Conf.*, **4**, 439 (1958); *Proc. Intern. Symp. Isotope Separation, Amsterdam*, (1956).
12. Kistemaker, J., Los, J., and Veldhuyzen, E. J., *Proc. Intern. Symp. Isotope Separation, Amsterdam*, (1956).
13. Cohen, K., "Theory of Isotope Separation," New York, McGraw-Hill Book Co., Inc., 1951.
14. Benedict, M., and Pigford, T. H., "Nuclear Chemical Engineering," New York, McGraw-Hill Book Co., Inc., 1957.

Cross-references: BROWNIAN MOTION, MOLECULAR WEIGHT, ROTATION—CIRCULAR MOTION.

CERENKOV RADIATION

This is a feeble radiation in the visible spectrum, which occurs when a fast charged particle traverses a dielectric medium at a velocity exceeding the velocity of light in the medium. It is thus a shock-wave phenomenon, the optical analog of

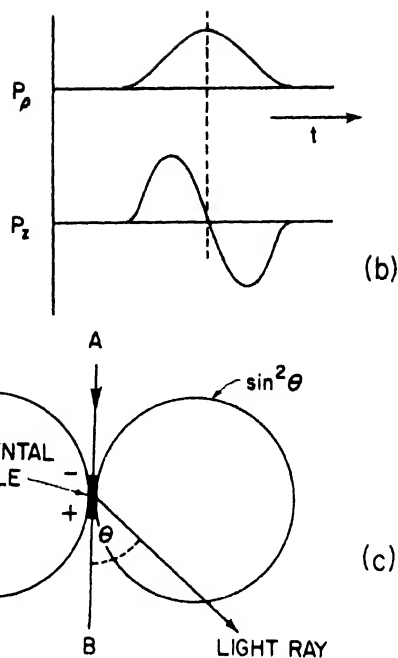
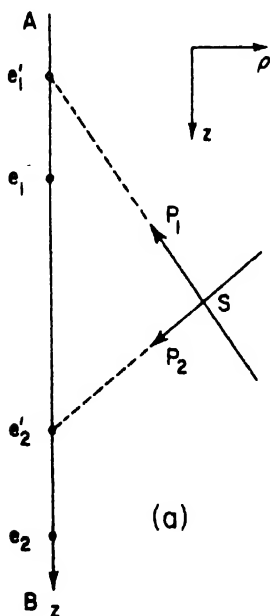


FIG. 1

the "supersonic bang." The radiation arises from the local and transient polarization of the medium close to the track of the particle. Consider, Fig. 1(a), an arbitrary element S of the medium to one side of the track AB of a fast electron, the track defining the z -axis. At a particular instant of time, when the electron is at say e_1 , the local polarization vector P will be directed along Se_1' , to a point e_1' slightly behind e_1 , owing to the retarded fields. As the particle goes by, the vector P will turn over and, when the electron reaches e_2 , will be directed to a point e_2' . The variation of P with time, may be resolved into radial and axial components P_ρ and P_z , as shown in Fig. 1(b). Owing to cylindrical symmetry, this polarization, viewed at a point distant from the particle, appears as an elementary dipole lying along the axis z , Fig. 1(c). As the particles plunge through the medium, radiation arises from the coherent growth and decay of this sequence of elementary dipoles. Two essential features of the radiation become at once apparent. First, since it is only the P_z component which is important, the field variation, Fig. 1(b), is that of a double δ -function. Thus, from Fourier analysis, if the circular frequency is ω , we will expect a spectrum of the form $\omega^2 d\omega$, i.e., radiation which is bluer than that from an equi-energy spectrum. Secondly, since the radiating element is an axial dipole, the angular distribution, for this element alone, will be of the form $\sin^2 \theta$, Fig. 1(c). It is important to realize that the radiation arises from the medium itself, not directly from the particle. Since the medium is stationary, the intensity and angular distribution do not contain the relativistic factor (mc^2/E) ;



in this respect, it is essentially different from Bremsstrahlung or synchrotron radiation.

The description above applies only to one element along the track. The most characteristic feature of Čerenkov radiation, its coherence, is at once apparent when we now consider an extended region of track. In Fig. 2(a) it is easily seen that there is only one angle θ at which it is possible to obtain a coherent wave front. If the velocity of the particle is $v = \beta c$, where c is the velocity of light in vacuo, and n is the refractive index of the medium, the particle travels a distance AB in a time Δt , given by $AB = \beta c \cdot \Delta t$; in the same time the radiation, emitted at A ,

travels a distance $AC = (c/n)\Delta t$, from which we obtain the Čerenkov relation:

$$\cos \theta = (1/\beta n) \quad (1)$$

From Eq. (1) it is at once evident that there is a threshold velocity given by $\beta = (1/n)$, below which no radiation takes place. At ultrarelativistic velocities, as $\beta \rightarrow 1$, the Čerenkov angle θ tends to a maximum value $\theta(\max) = \cos^{-1}(1/n)$. The polarization vectors E and H of the radiation which, owing to symmetry takes place over the surface of a cone, are shown in Fig. 2(b).

The radiation yield, from the theory of Frank and Tamm, is

$$\frac{dW}{dz} = \frac{e^2}{c^2} \int_{\beta n > 1} \left[1 - \frac{1}{\beta^2 n^2} \right] \omega \cdot d\omega, \text{ ergs/cm path}$$

or (2a)

$$\frac{dN}{dz} = 2\pi \left(\frac{e^2}{hc} \right) \cdot \left[\frac{1}{\lambda_2} - \frac{1}{\lambda_1} \right] \cdot \sin^2 \theta \text{ photons/cm path}$$

(2b)

between wavelength limits λ_1 and λ_2 (in cm). The spectral distribution is $(dW/d\omega) \propto \omega$ or $(dW/d\lambda) \propto \lambda^{-3}$, expressed as energy per unit circular frequency or per unit wavelength, respectively. The radiation has, therefore, a continuous spectrum toward the blue and ultraviolet. There is no radiation in the x-ray region, for which $n < 1$. For example, in the case of a fast electron in water, $n = 1.33$, we find from Eq. (2b), that when $\beta \rightarrow 1$ and $\theta(\max) = 41^\circ$, the yield (dN/dz) is ~ 200 photons/cm, between λ_1 and λ_2 of 3500 and 5500 Å, respectively.

The phenomenon has found considerable application in the fields of high-energy nuclear physics and cosmic-ray research; in almost all practical Čerenkov counters, the light is detected by means of a photomultiplier. The unique directional and threshold properties of the radiation may be used in a number of different ways. For example, by velocity selection, it is possible to distinguish between particles of different mass having the same energy, and it is also possible to measure particle velocities directly, by measuring θ . Other examples may be cited: The e^2 dependence, Eq. (2) above, has been used to determine the charge spectrum of the primary cosmic rays, and transparent lead-loaded glasses have been developed as total-absorption spectrometers for high-energy γ -rays.

Čerenkov radiation in gaseous media is now used extensively in high-energy physics, light flashes from the night-sky associated with cosmic ray showers have been attributed to the effect, and microwaves have been produced by the Čerenkov process.

J. V. JELLEY

References

- Čerenkov, P. A., *C.R. Acad. Sci. (USSR)*, **2**, 451 (1934).
 Frank, I. M., and Tamm, Ig, *C.R. Acad. Sci. (USSR)*, **14**, (3), 109 (1937).

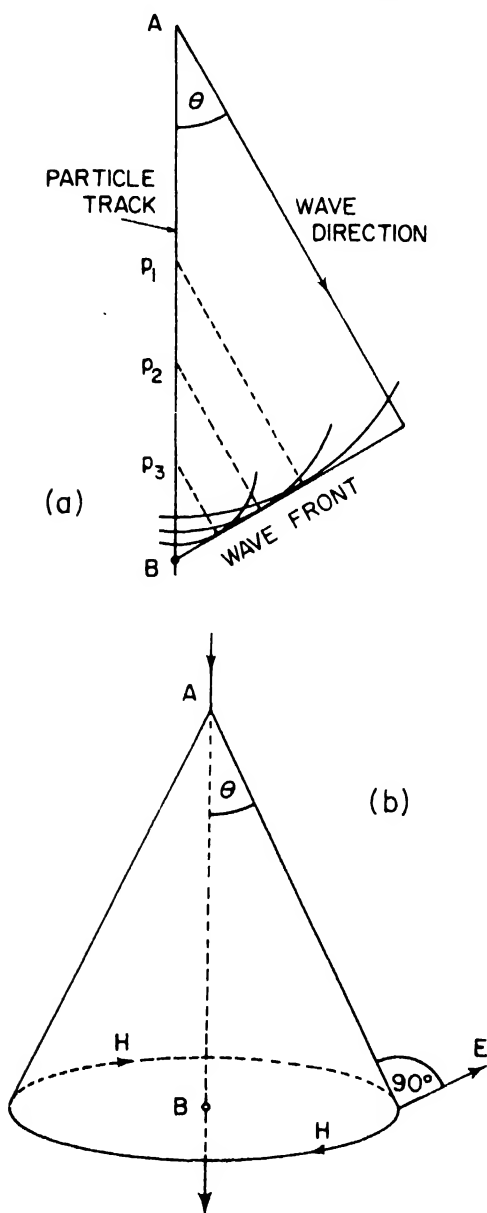


FIG. 2

Jelley, J. V., "Čerenkov Radiation and Its Applications," London, Pergamon Press, 1958.
 Bolotovskii, B. M., *Usp. Fiz. Nauk.*, **67**, 201 (1957).
 Hutchinson, G. W., *Progr. Nucl. Phys.*, **8**, 197 (1960).

Cross-references: DIELECTRIC THEORY, DIPOLE MOMENTS.

CHEMICAL KINETICS

Besides offering useful rate equations to describe the speeds of chemical reactions, chemical kinetics attempts to describe exactly how each reaction occurs. It does so in terms of one or more elementary steps, which are reactions having no observable intermediate chemical species. The ultimate goal is a theory interrelating energy, structure, and time for these single chemical events. Many of the ideas developed since 1850, when the first quantitative rate study was made, have been extended by analogy to explain electron-hole processes in semiconductors, various solid-state processes, and thermonuclear reactions.

Reaction rates depend on the nature of the reactants, temperature, pressure, kind and intensity of radiation, nature of catalyst or solvent, and many other factors. The extent of a reaction can be followed by withdrawal of samples for early chemical analysis. It is more common, however, to analyze the main reaction mixture continuously and nondestructively by spectroscopic means or by observing physical properties like density, electrical conductivity, optical activity, dielectric constant, and so on.

The rate r of a reaction $aA + bB \rightarrow eE + fF$ is best related to the rate of change of the concentrations $[A]$ and $[B]$ of reactants by rewriting the equation as $0 = eE + fF - aA - bB$. If the change is an elementary step, its rate is

$$r = -\frac{1}{a} \frac{d[A]}{dt} = -\frac{1}{b} \frac{d[B]}{dt} = \frac{1}{e} \frac{d[E]}{dt} = \frac{1}{f} \frac{d[F]}{dt} = k[A]^m[B]^n$$

Here k is a rate constant independent of concentration. If the step is not elementary, its rate is often presumed to be expressible as proportional to certain empirically observed powers m, n, \dots of the concentrations of species present in the reaction vessel. The values of m, n, \dots need not be integers and cannot be predicted from the balanced chemical equation if the step is not elementary. The over-all order of a reaction is the sum of these exponents, and the order with respect to a particular species is its own exponent. The order of a reaction can be determined in several ways. In the method of initial rates, concentrations of all but one reactant are held constant, if possible at great values, and v is observed at the start of reaction for several values of $[A]$. The order with respect to A is the slope of a graph of $\log v$ vs. $\log [A]$.

The first major reaction rate theory was founded on the kinetic theory and classical mechanics. It still is very useful when reactants

approach each other in an attractive potential field (e.g., ions), when the distribution of energies is nonequilibrium (e.g., electrical discharge in gases), or when the molecules involved are tremendous in size. In general, however, the collision theory suffers from its inability to satisfactorily predict effective cross sections (molecular sizes for reaction) or the effect of isotopic substitution.

Most modern theories suppose the existence of an undetectable transition state of high energy and fleeting existence. The configuration of this activated complex lies, as it were, atop the mountain pass of lowest height between energy-valleys of reactants and products. It is mechanically stable to all vibrations except the one that describes the progress of reaction over the saddle-point at the pass. This one motion is assigned a very low (sometimes imaginary) frequency, but otherwise the activated complex is just another molecule. It is supposed to be in dynamic equilibrium with reactants and its free energy can be calculated from its partition function by the usual methods of STATISTICAL MECHANICS. The rate constant k , then takes the form $(kT/h) \exp(-\Delta A/kT)$ where k is the gas constant per molecule, T is the absolute temperature, h is Planck's constant, and ΔA is the increase in free energy on going from reactants to activated state. This resembles the well-known relation $k = A \exp(-E/kT)$ discovered empirically in 1889 by Arrhenius. In it, A and E are approximately independent of T , and E is called the activation energy.

A recent theory of promise treats reactants as a wave packet that gradually spreads in time. The rate of reaction is taken to be the probability that the packet will be found in a configuration that is identified with products.

Thermal decomposition of an initially pure gas is seemingly a simple change, yet its order often changes gradually from first to second as the pressure falls. Moreover, there is always the question why like molecules do not all decay at once. The answers lie in understanding the mechanism, which is presently taken to be collision with any other molecule M ($0 \rightarrow A^* + M \rightarrow M + A$ with rate constant k_2') to yield an energized molecule A^* that may suffer a stabilizing collision ($0 \rightarrow A + M \rightarrow M + A^*$, with k_2) or internal change ($0 \rightarrow A^\ddagger \rightarrow A^*$, with k_1) that leads to the activated state A^\ddagger . The reaction is called unimolecular because the activated complex A^\ddagger contains only one reactant molecule. The rate of decomposition of A at any instant is $v = k_2'[M][A] - k_2[M][A^*]$, and the rate of decomposition of A^* is $v^* = k_1[A^*] - k_2'[M][A] + k_2[M][A^*]$. Since the v 's are time derivatives, these are simultaneous differential equations.

It is generally impossible to solve the simultaneous differential equations that describe a mechanism. The least restrictive and most useful simplification is generally the steady-state approximation, wherein the concentration of a species of low concentration is assumed to reach an effectively constant value after a certain

reasonable time (induction period) has passed. If $[A^*]$ reaches a steady-state concentration, $v^* = 0$ and the v^* differential equation becomes an algebraic one for $[A^*]$. Moreover, the rate of decomposition of A then is $v = k_1[A^*] = k_1k_2[M][A](k_1 + k_2[M])^{-1}$. This rate equation becomes second order if $k_2[M] \ll k_1$; this occurs at low pressure or when A^* changes rapidly into A^\ddagger . If A is a simple molecule of few atoms, the activation energy easily becomes effective in one bond to cause decay. On the other hand, the order becomes first if $k_2[M] \gg k_1$; this corresponds to high pressure or an A with many degrees of freedom to accommodate the activation energy.

A generally more restrictive way to simplify the mathematics of a sequence of reactions is to assume that one step is so much slower than the others that it alone limits the rate. All steps besides the rate-limiting one are assumed to be at equilibrium in this approximation. If, in unimolecular decomposition, the rate-limiting step is $A^* \rightarrow A^\ddagger$, then the rate is $k_1[A^*]$. The reaction is then first order in A because $v = 0$ for the equilibrium $A + M \rightleftharpoons M + A^*$. If, however, the rate-limiting step is $A + M \rightarrow A^* + M$, then the rate is $k_2[M][A]$.

Bimolecular reactions, wherein the activated state consists of two reactant species, are very common. The rates of the fastest of these are limited by the rates of diffusion of reactants and have rate constants of the order of 10^{10} liter mole⁻¹sec⁻¹ in aqueous solution. Typical examples are the aqueous neutralizations $NH_4^+ + H_2O \rightarrow NH_3 + H_3O^+$ and $H_3O^+ + F^- \rightarrow HF + H_2O$. Typical bimolecular gaseous reactions are the linear chain reactions $X + H_2 \rightarrow HX + H$ and $H + X_2 \rightarrow HX + X$ where X is H, D, Cl, Br, or I. A nice way to initiate these reactions of H_2 and X_2 is by a photon: $X_2 + \text{photon} \rightarrow X + X$. Thermal dissociation of X_2 is also sufficient to start reaction.

Carbon compounds undergo many reactions, but most of them can be classified into a few types. Nucleophilic substitution, wherein a basic reactant replaces another initially on C by a net reaction $X + RY \rightarrow XR + Y$, may be first order in RY alone or in both X and RY. If first order in just RY, the mechanism is labeled S_N1 and the rate-limiting step is conceived as production of the active carbonium ion R^+ by the process $RY \rightarrow R^+ + Y^-$. If second order (S_N2), the rate-limiting step is considered to be production of the bimolecular activated complex $X \cdots R \cdots Y$. Elimination reactions typically yield a double bond with loss of part of the organic reactant RCH_2CH_2Y . If first order in organic reactant (type E_1), the rate-limiting step is said to be production of the carbonium ion $RCH_2CH_2^+$, which then swiftly eliminates H^- to become $RCH = CH_2$. If first order in both base and organic reactant (type E_2), the rate-limiting step is thought of as production of a bimolecular complex which eliminates H^+ and Y^- almost simultaneously. A fifth class of organic reaction (S_N1) describes how an electrophilic reagent (e.g., NO_2^+ in mixed HNO_3 and H_2SO_4 or Br^+ in

Br_2 with $FeBr_3$) may attack an aromatic ring like that in benzene to form a positively charged intermediate that soon loses H^+ to a base in the solution.

A catalyst is a species that changes the rate of a reaction and yet is regenerated by that reaction so that it seems to be unchanged in the net reaction. Catalysts do not affect the equilibrium state but they do lower the activation energy and sometimes may provide a needed steric arrangement. The most general mechanism of catalysis is $A + C_1 \rightarrow D + C_2$ followed by $B + C_2 \rightarrow E + C_1$ to give the net change $A + B \rightarrow D + E$. Many so-called catalysts of industry need regeneration ($C_2 \rightarrow C_1$) by a reaction other than that catalyzed. For example, silica-alumina cracking catalysts used in making gasoline must be cyclically burned free of carbon deposited during cracking. A catalyzed reaction is almost always first order in catalyst concentration (or surface area) and usually has an order that is less than the true order by unity.

Enzymes are biological catalysts. Many act by the well-known Michaelis-Menten mechanism $E + S \rightleftharpoons C \rightarrow P + E$, where enzyme E attacks substrate S with a rate $k_A[E][S]$ to form a complex C that may yield products P at a rate $k_P[C]$ or may revert to S with rate $k_R[C]$. The rate of disappearance of S is $v_S = k_A[E][S] - k_R[C]$ and the rate of appearance of C or disappearance of E is $v_E = k_A[E][S] - k_R[C] - k_P[C]$. The total concentration of enzyme in the system is $[E]_0 = [C] + [E]$. In the steady state, $v_E = 0 = k_A([E]_0 - [C])[S] - (k_R + k_P)[C]$ so that

$$v_S - k_P[C] = \frac{k_A k_A[S][E]_0}{(k_R + k_P) + k_A[S]}$$

The rate of disappearance of S is always first order in total enzyme but may change from first to zero order in S as $[S]$ increases. The maximum rate at which E can act occurs when $[S]$ is great and $v_S \rightarrow k_P[E]_0$.

WILLIAM F. SHEEHAN

References

- Most physical chemistry textbooks contain introductions to chemical kinetics. Some of their many authors are: G. M. Barrow, F. Daniels and R. A. Alberty, S. Glasstone, E. A. Moelwyn-Hughes, W. J. Moore, and W. F. Sheehan. Benson, S. W., "The Foundations of Chemical Kinetics," New York, McGraw-Hill Book Co., Inc., 1960. Glasstone, S., Laidler, K. J., and Eyring, H., "The Theory of Rate Processes," New York, McGraw-Hill Book Co., Inc., 1941. Hinshelwood, C. N., "The Kinetics of Chemical Change," London, Oxford University Press, 1940. Slater, N. B., "Theory of Unimolecular Reactions," Ithaca, N.Y., Cornell University Press, 1959. There are several timely review articles in Volumes I, III, V, and VI of "Advances in Chemical Physics," I. Prigogine, Ed., London and New York, John Wiley & Sons, Ltd. (1958-1964).

Cross-references: CHEMISTRY, PHYSICAL CHEMISTRY, STATISTICAL MECHANICS.

CHEMICAL PHYSICS

It is difficult to define each of the traditional fields in science. It is more difficult to define in-between fields. This article will consist mostly of a description of chemical physics that follows brief historical descriptions of chemistry and physics.

During the nineteenth century chemistry dealt with atoms and molecules. It was the science of atoms and their combination. Chemistry was largely empirical and descriptive. Chemists devoted great effort to chemical synthesis and analysis. They made use of precipitation, titration, color changes, weighing, vapor pressure measurements, melting point measurements, and the like. On the other hand, physics dealt with mechanics, electricity and magnetism, wave motion, large-scale properties of matter and so forth. Its tools were those for measuring forces, velocities, electric fields, etc. Physicists had not learned how to treat small-scale mechanics. During this period chemistry and physics seemed to be far apart.

During the early part of the twentieth century physical science leaped ahead. Chemists began to make general use of classical thermodynamics and they became more interested in microscopic phenomena. By 1927 the basic structure of quantum mechanics had been erected and subsequently statistical mechanics was correspondingly modified and expanded.

By 1930 the separation between chemistry and physics no longer existed. The advent of statistical and quantum mechanics and of new, mostly physical, experimental methods and tools wrought great changes in chemistry and physics. The two sciences were able to help each other greatly, and the distinction between them became largely meaningless. It is perhaps unfortunate that chemistry and physics were ever separated. Surely it is true today that chemists need to know much about "physics" and vice versa.

There are many physical problems in which the concepts and methods of the two separate sciences are mixed so that this area may be called chemical physics. A wide range of study is common to both chemistry and physics. The basic problems in chemical physics concern properties of atoms and molecules and the behavior of statistical ensembles of atoms and molecules. Chemical physics is the study of the detailed spatial structures of atoms and molecules, the properties of matter on an atomic and molecular scale, and the application of results to macroscopic properties of matter. Physical chemistry, another interdisciplinary field, is more often concerned with the physical and thermodynamic properties of matter in bulk. Physical chemistry is usually regarded as a branch of chemistry.

We may use *The Journal of Chemical Physics* to explore chemical physics further. The fact that research men chose this journal for publication and that the editors of the journal did publish their papers would seem to indicate that the articles appearing in them "operationally define" chemical physics. Volume 40 of *The Journal of Chemical Physics* was issued during the first half

of 1964. During this interval 529 papers and 189 notes appeared.

The Journal of Chemical Physics lists 35 areas within chemical physics that appeared five times or more. Twelve per cent of these papers were on the kinetics of chemical reactions while 7 per cent dealt with the molecular structure. Within the top 50 per cent there were also, in descending order, papers on crystalline state, thermodynamic properties of matter, liquids, isotope effects, radiation effects, radiation chemistry, and photochemistry. Other topics, in descending order, were solutions, crystal structure, surfaces, energy transfer, polymers, viscosity, diffusion, electrochemistry, free radicals, dielectrics, susceptibility, transport phenomena, valence and chemical bond, rotation within matter, relaxation processes, gases, molecular interactions, phase transitions, conductance, dipole moments, electrical properties, flames, photoconductivity, semiconductors, accommodation coefficients, and photoionization. Each of these topics might well be listed in one or several journals devoted to chemistry, physics, physical chemistry, solid-state physics, etc. It is this combination of topics that is characteristic of the area of chemical physics.

The Journal of Chemical Physics indicates 20 tools or methods that were used by chemical physicists in carrying on their research. Twenty-two per cent used absorption spectroscopy (infrared, visible, and ultraviolet); 18 per cent used magnetic resonance. Within the top 50 per cent, in descending order, were: mass spectroscopy, emission spectroscopy, luminescence, and scattering (molecular, atomic, and x-ray). Also employed were: irradiation, beams (atomic and molecular), x-ray diffraction, conductance, absorption spectroscopy (microwave), Raman spectroscopy, shock waves, nuclear quadrupole resonance, magnetic susceptibility methods, high pressure, light scattering, neutron diffraction, ultrasonics, and electron diffraction. One should note that all of these methods have been developed since 1910. Five have been developed since 1945.

Theoretical methods were identified for 200 papers. They were quantum mechanics (67 per cent), statistical mechanics (26 per cent), thermodynamics (4 per cent), and classical mechanics (3 per cent). Theoretical analyses of microscopic systems and phenomena play a major role in chemical physics. About 40 per cent of the papers were primarily theoretical.

The author knows of eleven graduate schools in the United States that offer majors in chemical physics. It is estimated that the number is in excess of twenty. In many instances these programs are specifically set up for chemical physics students. In other cases the student is simply guided in his selection of chemistry, physics, and mathematics courses.

Other articles in this book discuss in detail some of the major areas in chemical physics.

J. W. McGRATH

Cross-references: CHEMISTRY; MATHEMATICAL PHYSICS; PHYSICAL CHEMISTRY; PHYSICS; THEORETICAL PHYSICS.

CHEMISTRY

Chemistry is the branch of natural science that includes knowledge about the nature, composition and transformation of matter and the particular structure of substances and compounds. Because every chemical change always involves one or more physical interactions, chemistry is very closely allied to physics. In fact, current knowledge of atomic structure, both nuclear and extranuclear, was derived almost entirely by using physical concepts and techniques.

Material creativity is dramatically manifested in chemistry, because in this field only is man able to synthesize new combinations of elements, substances and materials. Actually, chemists have created in the laboratory several chemical elements (43, 61, 87, and 93 to 103 inclusive) that apparently are not normally present in or on our planet. Even a cursory examination of the concept of isomerism (the existence of two or more different compounds that have identical compositions) yields the conclusion that literally trillions and trillions of different compounds of carbon could be synthesized by man. In fact, there are possible at least $62 \cdot 10^{12}$ different compounds (called isomers) all having an identical composition indicated by the formula $C_{10}H_{82}$. This specific combination of atoms is just one of millions of other possible combinations, and some of these would have thousands of isomers. The current chemical literature contains documented evidence for either the existence, or the synthesis by man, of well over a million different compounds of carbon. And carbon is only one of the 103 known chemical elements. No wonder, therefore, that chemical nomenclature and notation is of more than a little concern to modern chemists.

Because of the magnitude and latitude of chemistry, several specialized branches have arisen over the years. Some of these divisions are inorganic, organic, analytical, physical, physical organic, chemical physical, quantum chemistry, food chemistry, geochemistry and astrochemistry.

The chemical industry is now a significant part of our economy. The abundance of coal, petroleum and natural gas is one prominent reason for the growth of this branch of commerce. The continually increasing need for commodities such as structural materials, fabrics, fertilizers, pesticides and pharmaceuticals will enhance the chemical industry. The consumption of sulfuric acid, for instance, has for several decades been considered a significant indication of general industrial activity.

Historical Background. The history of alchemy is a fascinating record of man's earliest investigations of matter. However, the alchemist's productivity was seriously hampered by absence of such factors as logical reasoning, unbiased observations and interlaboratory communications. The presence of greed, in some cases, was detrimental.

Robert Boyle and Antoine Lavoisier, two of the earliest proponents of exact quantitative experimentation, exerted profound influences on late

eighteenth and early nineteenth century chemistry. Their techniques were mainly physical.

Around the middle of the nineteenth century, Friedrich Kekulé and Archibald Couper independently proposed a system for writing graphic formulas for chemical compounds. Their concepts were based apparently on the notion that physical forces held together the atoms in compounds. About the same time, general acceptance was finally accorded Amedeo Avogadro's hypothesis (proposed about 50 years earlier) which stated that equal volumes of gases at the same temperature and pressure contain equal numbers of molecules. Acceptance of this physical concept and those of Kekulé and Couper was a marked stimulus to nineteenth century chemical research.

During the period 1860–1900 organic chemistry flourished, and hundreds of compounds were prepared, correctly analyzed and identified, and logically classified as to structure and reactivity. And yet, throughout this period, chemists had no knowledge whatsoever about the structure and composition of atoms. Also they had a rather shallow conception of chemical bonding and molecular geometry. The growth of organic chemistry during the nineteenth century is an epic example of the productivity of sound inductive and deductive reasoning.

Some discoveries in physics that greatly accelerated the growth of chemistry are: Henri Becquerel's discovery of radioactivity (1896) which ultimately led to our current knowledge of atomic nuclear composition and phenomena; the belief in the significance of the electron (Stoney, 1894) which led to the acceptance of the existence of ions; Max Planck's quantum concept (1900); the concepts of probability and the equivalence of mass and energy by Albert Einstein (1905); Niels Bohr's atomic theory (1913); and the duality of the wave and particle character of matter by Louis DeBroglie (1924).

Since 1940 there has been phenomenal growth in the use of spectroscopy in the elucidation of the structure of molecules. The major types of spectra are, in order of seniority, ultraviolet, infrared, nuclear magnetic resonance, and electron spin resonance. Mass spectrometry is valuable in the determination of molecular as well as atomic structure.

During and following World War II there was considerable investigation of the transuranic elements, and of technetium, francium and promethium in relation to their production and use. All these radioactive elements were produced by extraordinary nuclear reactions, whereas all ordinary chemical reactions involve only the electrons outside the nucleus. Ordinary chemical reactions are, in a sense, extranuclear.

The extension, since 1940, of the techniques of column chromatography to other types such as paper, vapor phase (gas), salting-out, and thin-layer has made much easier the separation of the components of complex mixtures. The use of ion-exchange in the resolution of ionic mixtures and in the removal of ions from solutions has been extended during the past twenty years.

The applications of the molecular orbital, crystal-field and ligand-field concepts have been highly successful in the description of the bonding in and geometry of complex molecules. These concepts have firm physical bases.

The decision made by chemists and physicists in 1961 to use a common standard, carbon-12, to assign atomic masses to the chemical elements is a notable landmark in the history of chemistry.

The discovery in 1962 by Neil Bartlett of a stable compound of xenon, $\text{Xe}(\text{PtF}_6)$, was an epic event. Since the discovery of the noble gases during the period, 1894-1900, most chemists had assumed or believed that these elements were either chemically inert or could form only unusual compounds or complexes. Therefore, Bartlett's discovery prompted much activity to produce other noble gas compounds. Several stable binary fluorides, such as XeF_2 , XeF_4 , XeF_6 and KrF_4 are ordinary solid compounds produced by ordinary chemical reactions. Xenon oxy-compounds such as XeO_3 , XeOF_4 and XeOF_6 , although less stable than the binary fluorides, seem to be ordinary compounds.

The importance of chemistry in modern biology, and especially in medicine, has been re-emphasized by knowledge of the role of the nucleic acids, such as DNA and RNA, in the genetic scheme. The increased use of chemotherapy in medicine is common knowledge.

Inorganic Chemistry. This branch of chemistry is mainly that of all forms of noncarbonaceous matter. Although its potential scope is huge, it has attracted, until recently, much less attention than have organic and physical chemistry. The major minerals (except coal, petroleum and lignite) are essentially inorganic. The use of inorganic compounds, especially those of boron and of nitrogen, as fuels and the growth of non-ferrous (other than iron) metallurgy are major developments in inorganic chemistry.

Organic Chemistry. This division of chemistry is essentially that of the compounds of carbon. Originally organic chemistry was confined to materials in or from living organisms, probably because nearly every compound either isolated from or produced by a living organism is a compound of carbon. Although there are more compounds of hydrogen than of any other element, the compounds of carbon are next in line. The property of catenation (ability of identical atoms to bond together) is exhibited most extensively by carbon. The hundreds of different carbon-atom skeletons of the thousands of known organic compounds attest this fact. All foods, nearly every fabric, every ordinary commercial fuel, and almost all pharmaceuticals are organic in the sense they contain compounds of carbon.

Analytical Chemistry. The qualitative and quantitative determination of the elemental composition of matter resides in the branch of chemistry called analytical chemistry. Any means of determining molecular structure is often called an analytical technique. Until relatively recently most analyses were performed by using specific chemical reactions and techniques in liquid

solutions. Recent advances in spectroscopy and other physical techniques have yielded a variety of instruments that greatly facilitate chemical analyses.

Physical Chemistry. The quantitative measurements of the properties and behavior of the elements and their compounds are the major concern of the physical chemist. Nearly every technique and concept has been adopted from physics. The development of new chemical concepts follows logical consideration of quantitative data.

The major branches of physical chemistry are spectroscopy, nuclear chemistry, kinetics, thermodynamics, quantum and statistical mechanics, and solution and surface chemistry. Physical organic and inorganic chemistry have gained prominence during the past 25 years.

Biochemistry. Investigations of the chemical phenomena in, and the constituent compounds of, living organisms are performed mainly by biochemists. Because every chemical reaction in any living organism involves compounds of carbon, biochemistry is essentially the application of organic chemistry to investigations of vital systems. However, both physical and analytical chemistry are essential to biochemistry.

DONALD C. GREGG

Cross-references: BOND; CHEMICAL; ELEMENTS; ISOTOPES; MOLECULAR WEIGHT; MOLECULES; PHYSICAL CHEMISTRY; SPECTROSCOPY

CIRCUITRY

Basic Concepts. As with very many concepts relating to electricity and magnetism, that of the electric circuit very properly may be attributed to James Clerk Maxwell. It was he who took the bold step of ascribing a dual role to the quantity I , which he identified with the current. In perfect conductors, it is the rate of flow of electric charge. In perfect dielectrics, it is proportional to the time rate of change of electric intensity. Above all, however, it has the typical property that it always flows in closed paths. It is this last property upon which the whole concept of electric circuitry is based.

To be more explicit, consider an elementary circuit composed of a resistor R , a capacitor C , and an inductor L , all three connected together end-to-end so that a closed ring or loop is formed. Open this ring at some point and insert a source of electric energy, such as a battery or an alternator that causes an electric current to flow around the closed loop. At any point in the loop the value of the current is equal to its value at any other point. It is this property of the current that determines the configuration of the circuit. The relation of the current to the electromotive force of voltage V generated by the energy source is given by the familiar differential equation

$$V = RI + L \frac{dI}{dt} + \frac{1}{C} \int I dt \quad (1)$$

In a very general sense V may be expressed in the form of the sum of a number of exponentials

$$V = \sum b_m e^{p_m t} \quad (2)$$

where at least one of the b_m 's may be zero. This suggests trying to find solutions of Eq. (1) by expressing the current in the analogous form

$$I = \sum a_m e^{p_m t} \quad (3)$$

When Eqs. (2) and (3) are substituted in Eq. (1) it is seen that the p_m 's must be paired with identical p_m 's. This gives a set of equations each having the form:

$$b_m = R a_m + L p_m a_m + \frac{a_m}{C p_m} \quad (4)$$

There will be one such equation for each value of b_m in Eq. (2). They can be solved to give

$$a_m = \frac{b_m}{R + L p_m + \frac{1}{C p_m}} \quad (5)$$

As remarked above, at least one of the b_m 's may be zero. For this case, the denominator of Eq. (5) must also be zero. Thus, in Eq. (5) there are two values of p_m which apply, namely

$$p_{01} = \frac{R}{2L} \pm j \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}} \approx j\beta \quad (6)$$

$$p_{02} = \frac{R}{2L} \pm j \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}} \approx j\beta \quad (7)$$

When these are reintroduced into Eq. (2), the corresponding values of the a_m 's are determined by the known conditions existing at a given time, for instance at $t = 0$.

From this approach it is easy to generalize. A periodic voltage is the sum of two exponentials $b_1 e^{j\omega t}$ and $b_1 e^{-j\omega t}$. The resulting current is correspondingly the sum of two exponentials

$$\frac{b_1 e^{j\omega t}}{R + j\omega L + \frac{1}{j\omega C}} \quad \text{and} \quad \frac{b_1 e^{-j\omega t}}{R - j\omega L + \frac{1}{j\omega C}}$$

Since the second of these is merely the complex conjugate of the first, both for the voltage and the current, it is convenient for circuit analysis to deal only with the first in both cases. However, the convention breaks down in handling nonlinear systems. There it is necessary to carry the conjugate terms throughout the analysis.

A further step in the generalizing process is to regard Eq. (2) as a Fourier series where the p_m 's are now the imaginaries $\pm j\omega_0$. The quantity $R + j\omega L + 1/j\omega C$ is called the impedance Z . Hence, in a very simple way, for each of the exponentials comprising V , we can express Eq. (1) in the form

$$V = IZ \quad (8)$$

whence

$$I = \frac{V}{Z} \quad (9)$$

The solution of the elementary single-loop circuit is thus reduced to that of a simple algebraic equation.¹

The extension to multi-loop networks follows immediately. It is based on the concept that currents always flow in closed loops and that the values of the loop currents are always the same at any two points in the loop. Figure 1 gives a

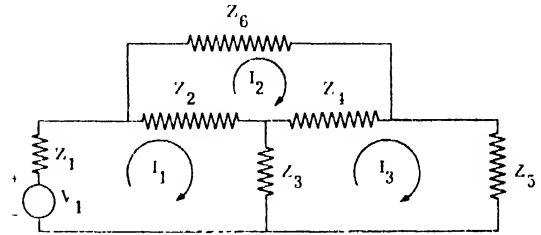


FIG. 1

typical configuration.* This shows a three-loop network where the loops are identified by the three currents, I_1 , I_2 and I_3 . In applying Eq. (8) to these three loops, it is necessary to take cognizance of the fact that some of the impedance elements, Z_2 for instance, are subjected to the influence of more than one of the individual loop currents. The current I_1 flows down through Z_2 and the current I_2 flows up through it. The net contribution of Z_2 to the right-hand side of Eq. (8) is $(I_1 - I_2)Z_2$. Thus, the complete description of the network of Fig. 1 is given by the three equations, one for each loop, as follows:

$$\begin{aligned} V_1 &= I_1(Z_1 + Z_2 + Z_3) - I_2 Z_2 - I_3 Z_3 \\ 0 &= -I_1 Z_2 + I_2(Z_2 + Z_4 + Z_6) - I_3 Z_4 \\ 0 &= -I_1 Z_3 - I_2 Z_4 + I_3(Z_3 + Z_4 + Z_5) \end{aligned} \quad (10)$$

Solution yields the currents in the form

$$I_n = \frac{V_{1n}}{Z_{1n}} \quad (11)$$

where Z_{1n} is the transfer impedance from V_1 to the n th current loop. Matrix methods of handling combinations of such networks immediately suggest themselves and these methods have been developed in considerable detail in the literature.^{2,3}

Another generalization involves extending Eq. (2) beyond the Fourier series form to the

* Certain complications occur when circuit networks are of such a topological character that they cannot be represented on a plane surface, as well as when magnetic couplings are involved. However, appropriate extensions of the simple concepts are available for handling these cases.

Fourier integral. Thus, instead of Eq. (2) we have

$$V(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} A(j\omega) e^{j\omega t} d\omega \quad (12)$$

where

$$A(j\omega) = \int_{-\infty}^{\infty} v(t) e^{-j\omega t} dt \quad (13)$$

In this context, Eq. (8) takes the form

$$I(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{A(j\omega)}{Z(j\omega)} e^{j\omega t} d\omega \quad (14)$$

and here all of the matrix generalizations indicated by Eq. (11) may be carried along. In this environment, many of the fundamental properties of the impedance function may be derived.⁴ For example, the fact that a real voltage produces only real currents insures that the real portion of the impedance function is an even function of ω , while the imaginary portion is an odd function. Consideration of the behavior of Eqs. (11), (12) and (13) over the entire complex plane of

$$p = \alpha + j\omega \quad (15)$$

yields further insight into the interesting and useful properties of the impedance function and the entire impedance concept in the environment of Fourier and Laplace transforms.^{4,5}

One of the more important of these expresses the current response of a network to any arbitrary voltage in terms of its response to a vanishingly short pulse of voltage. The response may be measured either in the same loop where the voltage is applied or in a different one. In this aspect, Eq. (11) takes the form

$$I(t) = \int_{-\infty}^{\infty} V(\tau) J(t - \tau) d\tau \quad (16)$$

where $J(t - \tau)$ is the current resulting from a very short pulse of voltage of unit amplitude injected at time $(t = \tau)$. The integral in Eq. (16) is known as the convolution of $V(\tau)$ and $J(t - \tau)$.

Other important properties derivable from consideration of the whole complex plane involve the properties of active vs passive networks, stability requirements for active networks, and requirements for physical realization imposed on the impedance elements and their properties.^{4,6,7}

Nonlinear Circuits. Solution of circuits involving nonlinear relations is in nowhere near as satisfactory a condition. This is not to deny that many of their properties have been discovered and analyzed in rather complete fashion. Those cases where the nonlinearity can be considered as a perturbation on the linear situation have been dealt with satisfactorily. Even in cases involving mixers and modulators, analysis and interpretation is quite straightforward.⁸ However, handling of such situations as the buildup and leveling off of transients in self-oscillatory circuits has made

little progress since the investigations of such pioneers as Balth Van der Pol.⁹

Systems With Time-varying Elements. A number of years ago, interest was aroused in the properties of circuits containing time-varying elements.¹⁰ It was then shown that some of these had amplifying properties. Since no very simple way of producing such elements developed at that time, interest lagged until a few years ago. Now it is realized that modern parametric amplifiers may be handled in this same manner. Moreover, many of the properties of masers and lasers may be interpreted in analogous terms.

Other New Circuit Elements. Of special interest in recent years are circuits that transmit signals differently in different directions. Of course, circuits with vacuum tubes had been known to have this property for many years. Under some conditions, circuits with transistors can be given the same property. These all require sources of power to realize their properties. Some configurations such as those of gyrators, isolators and circulators, possess the property without requiring such sources. True, they do demand the presence of a magnetic field but this need not necessarily require power consumption. The initial application of circuits containing isolators was to prevent changes in the neighborhood of a radio antenna from reacting backwards and affecting the properties of circuitry that was feeding signal energy into the antenna. Later applications allowed multistage parametric amplifiers to be built without encountering backward flow of signal energy in sufficient quantities to produce instabilities in the form of self-oscillations.

Physical Embodiments. While these advances and developments were taking place in the theoretical analysis and technical applications of circuitry, changes of far-reaching significance were in progress in their physical aspects and in materials composing the impedance elements themselves. Originally, resistors, inductors, and capacitors were distinct and separate entities of rather bulky form and unique appearance. Resistors required long lengths of german silver wire, wound so as to reduce the inductive effects, or else lengths of graphite or of volcanic material that went by the name of lavite. Inductors were coils of wire wound on an insulating form either with air core or a core of magnetic material such as soft iron. Capacitors were multi-layer affairs of metallic plates separated by dielectric material which might be either solid or air. Paper was often used for the dielectric.

Nowadays, complicated circuits consisting of many elements in different combinations are formed by various techniques of depositing, spraying or printing materials of many kinds on sheets or cards. Capacitors are formed with much higher capacitance values compressed into volumes much smaller than ever before. Similar advances have been made with inductors. Very thin films of metallic material have contributed to improvements in reliability, stability and accuracy of resistors also.

Switching circuitry has developed very rapidly

† Active networks are those that contain energy sources such as electron tubes or transistors.

in connection with electronic telephone exchanges and digital computers. Circuit elements involve many new devices such as ferreed switches, ferrite memory storage and twistors. Also, it should be noted that optical methods are increasingly being adapted to perform the functions of the older types of circuitry including even such applications as filtering and equalizing. It is probable that the next few years will see remarkable developments and applications along these lines as well as in other directions as yet not even thought of.

F. B. LLEWELLYN

References

1. Carson, J. R., "Electric Circuit Theory and Operational Calculus," New York, McGraw-Hill Book Co., 1926.
2. Guillemin, A. E., "Communication Networks," New York, John Wiley & Sons, 1931; "Theory of Linear Physical Systems," John Wiley & Sons, 1963.
3. Weinberg, Louis, "Network Analysis and Synthesis," New York, McGraw-Hill Book Co., 1962.
4. Bode, H. W., "Network Analysis and Feedback Amplifier Design," New York, John Wiley & Sons, 1945.
5. Brown, W. M., "Analysis of Linear Time-Invariant Systems," New York, McGraw-Hill Book Co., 1963.
6. Nyquist, H., "Regeneration Theory," *Bell System Tech. J.*, **11**, 126-147 (January 1932).
7. Llewellyn, F. B., "Some Fundamental Properties of Transmission Systems," *Proc. I.R.E.*, **40**, 271-283 (March 1952).
8. Llewellyn, F. B., and Peterson, L. C., "The Performance and Measurement of Mixers in Terms of Linear Network Theory," *Proc. I.R.E.*, **33**, No. 7, 458-476 (July 1945).
9. Van der Pol, B., "Selected Scientific Papers," Amsterdam, North-Holland Publishing Co., 1960.
10. Hartley, R. V. L., "Oscillations in Systems with Non-linear Reactance," *Bell System Tech. J.*, **15**, 424-440 (July 1936).

Cross-references: ALTERNATING CURRENTS; CAPACITANCE; CONDUCTIVITY, ELECTRICAL; FEEDBACK; INDUCTANCE; LASER, MASER.

CIRCULAR MOTION. See ROTATION—CIRCULAR MOTION.

CLOUD PHYSICS

Gross Dynamics of Clouds. Atmospheric clouds consist of small drops of water condensed from atmospheric water vapor by cooling of the air below its saturation temperature (dew point). Cooling may be induced by radiation near the surface or by temperature contrasts between the air and underlying surface to form clouds near the surface (fog.) Clouds in the upper atmosphere are mostly caused by more-or-less adiabatic cooling due to lifting of the air. Gradual upward vertical motions induced by large-scale atmospheric circulations (~100 km or more in diameter)

usually lead to stratiform clouds, whereas small-scale turbulent motions usually induced by heating from below may result in bunched dense clouds called cumulus (~100 meters to several kilometers in diameter and height).

In meteorological work, a thermodynamic chart with pressure and temperature as the ordinate and abscissa, respectively, is often used to relate the presence of clouds to the vertical temperature and humidity structure of the atmosphere at a given radiosonde station. On such a chart, the adiabatic behavior of an imaginary air parcel may be followed in an imaginary lifting process from any level, particularly the surface, to establish the condensation level (cloud base). The atmosphere's vertical stability may also be evaluated based on whether or not the parcel is negatively (stable) or positively buoyant (unstable) with respect to the atmospheric sounding.

A more realistic representation of convection in the atmosphere may be made by allowing mixing in proper proportions between the parcel and its environment. Observations with a properly instrumented aircraft may eventually be used to relate temperature, liquid water content, drop size and vertical velocity in cumulus clouds by a physically consistent turbulent mixing process between cloud and environment. The relationship of cumulus convection to large-scale atmospheric motions and of the smaller-scale mixing process to the microphysical aspects of cloud drop growth and coalescence are understood only in a rough qualitative sense.

Microphysics of Clouds. Clouds of liquid drops can be formed in a pure atmosphere free of dust or charged particles only by supersaturations of three or four times that over a flat water surface. So many particles are suspended in the earth's atmosphere that maximum initial supersaturations of the order of 1 per cent are all that is required to start the condensation process, which proceeds without appreciable supersaturation once the drops are past a critical size. Computations of the early stages of drop growth indicate that the initial number of drops is established by the initial upward velocity and the number of extremely active condensation nuclei available. Over the oceans the active nucleus spectrum is dominated by sea salt particles of the order of 1μ in diameter (giant salt nuclei).

As a cloud grows, its top may cool below freezing. The cloud drops do not freeze immediately on reaching a temperature of 0°C but tend to supercool to considerably lower temperatures before freezing. Just as condensation is induced with difficulty in pure air, similarly freezing of pure water is also helped by the presence of impurities.

Particles in the air itself may become nuclei for sublimation (condensation from the vapor to the ice phase). Similarly particles in the cloud-drops may induce freezing at certain temperatures of supercooling. The results of laboratory experiments indicate that large water samples or drops tend to freeze at higher supercooling temperatures than small drops. The size dependence of average

freezing temperature has been explained as a sampling effect based on the greater probability of having an active nucleus in a large volume.

Ice-forming nuclei producing ice crystals at temperatures warmer than -20°C have a concentration of the order of one per liter of air. Therefore cloud tops as a rule do not become glaciated until the temperature of -20°C is reached.

The Formation of Rain. A cloud consisting of drops less than 40μ in diameter is quite stable in the sense that they grow only by condensation, not by coalescence. Recent work on the aerodynamics of small droplets indicates that collision of droplets less than 36μ in diameter is impossible. The collision efficiency of clouds consisting of droplets greater than 40μ in diameter rises rapidly as the drop size and drop concentration increase.

The presence of relatively few ice crystals in a supercooled cloud causes it to become unstable regardless of the drop size, since ice crystals at the temperature of supercooled water have a lower vapor pressure. Thus the ice crystals grow at the expense of drops and eventually fall through the cloud growing by coalescence.

Cloud Seeding. Awareness of ice crystals as a source of cloud instability led to cloud seeding experiments. Supercooled stratiform clouds were seeded by dropping dry-ice pellets from an aircraft flying in it. The clouds after seeding were observed to develop holes shaped like the aircraft's path configuration from which light snow showers fell.

Silver iodide in the form of smoke was later introduced as a seeding agent, because its crystal structure was very similar to that of ice. The effectiveness of silver iodide as a sublimation nucleus has been demonstrated in laboratory experiments, but field experiments have not been quite as obviously convincing. Cases of seeding of individual cumulus clouds have resulted in apparent dramatic changes in development, such as explosive vertical growth of firm white cloud tops into fuzzy ice crystal clouds.

Other seeding agents have been tried without convincing demonstrations of their efficacy. For example, clouds above freezing have been seeded with water sprays in attempts to induce a coalescence chain reaction. The injected large drops were expected to grow to a maximum stable diameter of .9 cm and break up into a number of smaller drops, which in turn were expected to grow. The injection of salt particles into clouds to furnish large condensation nuclei in continental clouds has been tried.

Implications for Weather Modification. The possibility of increasing rainfall by cloud seeding was noted in the early stages of experiments on cloud modification. Unfortunately the relation of rainfall observed at the ground to the effect of cloud seeding has proved extremely difficult in the face of the extreme natural variability of precipitation. So far, careful statistical design of weather modification experiments has not produced positive results.

Other forms of weather modification have been

tried or considered. Hail suppression with silver iodide rockets has been attempted without proof of effectiveness. An attempt to weaken a hurricane by seeding with silver iodide smoke bombs has been made with inconclusive results. The use of black topping on the earth's surface is being considered as a means of stimulating cumulus convection by surface heating so as to increase rainfall in favorably situated arid regions. Perhaps a definitive answer to questions on the possibility of weather modification must await a better understanding of the atmosphere.

JOSEPH LEVINE

References

- Bartan, I. J., "Cloud Physics and Cloud Seeding," Garden City, N.Y., Doubleday & Co., Inc., 1962.
 Fletcher, N. H., "The Physics of Rainclouds," Cambridge, Cambridge University Press, 1962.
 Mason, B. J., "The Physics of Clouds," London, Oxford University Press, 1957.

Cross-references: CONDENSATION, METEOROLOGY, THERMODYNAMICS.

COHERENCE

Basic Definitions. The term coherence as it is used in electromagnetic radiation studies is best explained by a discussion of Young's interference experiment. Referring to Fig. 1, we consider a self-luminous radiating source S (like a mercury arc lamp) placed a distance l_1 from an opaque screen A . In the screen A two pinhole openings P_1 and P_2 are made a distance d apart. The radiation passing through the pinholes impinges upon and is recorded upon a photographic plate B a distance l_2 from screen A .

If l_1 is a very large distance from A (say $l_1 \gg bd/\lambda$, λ being the average radiation wavelength) and the spectral width of the radiation is made very narrow by filtering, then the fringes observed on B in the neighborhood of O will be very sharp, and we say that the radiation fields impinging upon P_1 and P_2 are very coherent. This is in accord with our usual notions about the radiation from a point source. If, on the other hand, we use the same source and l_1 is taken to be small (say of the order of b) and the dimension b is large compared to d then fringes will not be observed on B , and we say that the radiation fields impinging upon P_1 and P_2 are incoherent. Again this is to be expected since in this case we observe the radiation just as it emerges from the lamp. As the distance l_1 is increased from the order of b , faint fringes will begin to appear upon B . As l_1 increases, the fringes will become progressively stronger until they become very pronounced when $l_1 \gg bd/\lambda$. The intermediate states, when the fringes are present but not necessarily very strong, are termed states of partial coherence. This experiment may be performed using a variety of sources. The experiment, as we have emphasized, measures the coherence of the radiation when it reaches P_1 and P_2 ; it does not measure the coherence of the source.

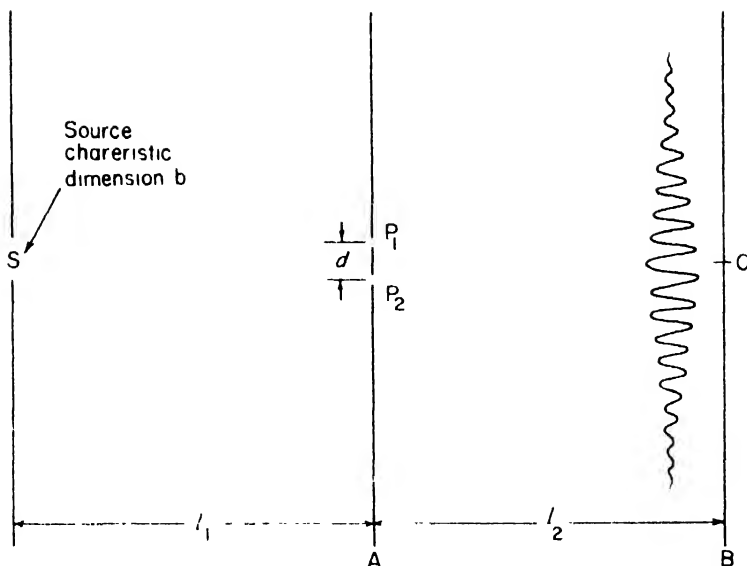


FIG. 1. Young's interference experiment.

A quantitative measure of the fringe strength called the visibility \mathcal{V} , was given by Michelson.¹ It is defined as

$$\mathcal{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \quad (1)$$

where I_{\max} is the maximum intensity recorded in the vicinity of O and I_{\min} is the minimum intensity. \mathcal{V} varies between 0 and 1 and, roughly, we term radiation fields at two points coherent when $\mathcal{V} \approx 1$ (strong fringes) and incoherent (no fringes) when $\mathcal{V} \approx 0$.

In the above introduction, d was held fixed but it is most important to realize that the fringe visibility at O is a function of the spacing d and that \mathcal{V} may be close to 1 for one spacing of d and close to zero for another spacing of d . For example, if we fix l_1 , we will generally decrease the visibility of fringes at O by making d larger (more precisely it will decrease and then increase in a succession of oscillations with the successive peaks being reduced in magnitude.) The visibility may thus be viewed as a measure of the correlation (coherence) between the radiation at P_1 and the radiation at P_2 . To give a more precise definition of coherence, Wolf² considered the correlation function $\Gamma(P_1, P_2, \tau)$, commonly termed the mutual coherence function, defined as

$$\Gamma(P_1, P_2, \tau) = \langle V(P_1, t + \tau) V^*(P_2, t) \rangle \quad (2)$$

where $V(P_1, t)$ is the radiation field at P_1 and time $t + \tau$ (in a scalar approximation), V^* is the radiation field at P_2 at time t and the brackets $\langle \rangle$ denote a time average. For convenience Wolf used a complex notation (V^* is the complex conjugate of V), but this need not concern us here.

It can be shown that the magnitude of the

normalized form of $\Gamma(P_1, P_2, \tau)$, $\gamma(P_1, P_2, \tau)$, where

$$\gamma(P_1, P_2, \tau) = \frac{\Gamma(P_1, P_2, \tau)}{\sqrt{\Gamma(P_1, P_1, 0)\Gamma(P_2, P_2, 0)}} \quad (3)$$

is equal to the visibility \mathcal{V} when the radiation is quasi-monochromatic (narrow spectral width) and all path lengths in the problem are small compared to $c \Delta\nu$ (where c is the velocity of light and $\Delta\nu$ is a characteristic spectral spread). In this case it is appropriate to fix τ at some value τ_0 usually taken to be zero.

The definition given by Eq. (2) was intended to consider polychromatic fields in addition to the quasi-monochromatic fields so often studied using the visibility \mathcal{V} . Beran and Parrent¹ have shown that the full function $\Gamma(P_1, P_2, \tau)$ is, in principle, measurable by an extension of the techniques used in a Young's interference experiment. Modern discussions of coherence now center on the calculation and measurement of the mutual coherence function $\Gamma(P_1, P_2, \tau)$. For calculation it is convenient to note that Wolf² has shown that $\Gamma(P_1, P_2, \tau)$ satisfies the pair of wave equations

$$\nabla_i^2 \Gamma(P_1, P_2, \tau) = \frac{1}{c^2} \frac{\partial^2 \Gamma(P_1, P_2, \tau)}{\partial \tau^2} \quad (i = 1, 2) \quad (4)$$

The vector properties of the electromagnetic field may easily be introduced into the mutual coherence function. In general one must consider the tensor

$$\epsilon_{ij}(P_1, P_2, \tau) = \langle E_i(P_1, t + \tau) E_j^*(P_2, t) \rangle \quad (5)$$

where $E_k(P_k, t)$ is the k th component of the electric field at the space point P_k and time t . In general, this function has proved most useful in the study of polarization effects for fields in

which the radiation has a principal direction of propagation. The form of $\delta_{ij}(P_1, P_2, \tau)$ has, however, been derived for the radiation in a black body cavity.

The above considerations have been for radiation fields that are stationary in time. That is, the statistical parameters, as opposed to the detailed structure of the radiation field, are independent of the absolute scale of time. For fields in which this is not true we must introduce the concept of an ensemble average (similar to that used in statistical mechanics). The mutual coherence function $\delta_{ij}(P_1, P_2, \tau)$ must be replaced by the function $\delta_{ij}^E(P_1, t_1; P_2, t_2)$ defined as

$$\delta_{ij}^E(P_1, t_1; P_2, t_2) = \overline{E_i(P_1, t_1)E_j(P_2, t_2)} \quad (6)$$

where the overbar denotes an average over an ensemble of systems.

Higher-order Coherence Functions and Quantum Aspects. As the concept of coherence grew and the statistical formulation of the theory came more into the fore it was realized that a two-point moment like $\delta_{ij}(P_1, P_2, \tau)$ or $\delta_{ij}^E(P_1, t_1; P_2, t_2)$ was inadequate to completely describe the radiation field. Two fields could have the same mutual coherence function and yet differ in the statistical content of the field. To completely describe the field, it was necessary to consider higher-order moments like $L_{ijkl}(P_1, t_1; P_2, t_2; P_3, t_3; P_4, t_4)$ defined as

$$L_{ijkl}(P_1, t_1; P_2, t_2; P_3, t_3; P_4, t_4) = \overline{E_i(P_1, t_1)E_j(P_2, t_2)E_k(P_3, t_3)E_l(P_4, t_4)} \quad (7)$$

and, in fact, for a complete description it was necessary to consider the probability density functional $P[E_i(P_1, t_1)]$ defined roughly as the probability of the occurrence of a particular realization of the field.

At the writing of this article there is no consensus in the field as to a definition of coherence in terms of all higher moments, and the definition of coherence in terms of second-order moments is still commonly used. There are, however, a number of problems that require consideration of higher-order moments. These moments are necessary in intensity interferometry, the study of laser radiation and the study of the radiation from turbulent gases.

The measurement of the contracted fourth-order moment

$$R_{ik}(P_1, P_2, 0) = \langle [E_i(P_1, t)]^2 [E_k(P_2, t)]^2 \rangle \quad (8)$$

(called intensity interferometry) has received considerable attention since it entailed consideration of the quantum aspects of the electromagnetic field. This moment may be thought of as the correlation of instantaneous intensities

$$I_i(P_1, t) \equiv [E_i(P_1, t)]^2$$

at two points and was originally studied as an alternate method to the measurement of $I'(P_1, P_2, 0)$ for determining the angular diameter of

visible and radio stars.^{4,5} In measuring $I'(P_1, P_2, 0)$, we may use the Young's interference experiment described above; the quantum nature of the field rarely explicitly enters since averaging times are usually long enough to permit a classical analysis. To measure R_{ik} , however, we need to correlate the two signals $I_i(P_1, t)$ and $I_k(P_2, t)$ which are recorded using photomultipliers or coincidence counters. Since the relationship between the impinging electric field and the ejected photoelectron is a statistical one, classical considerations did not suffice for a deep understanding of the problem or for consideration of the very important signal-to-noise problems resulting from inadequate averaging times.

Many studies of the quantum aspects of coherence theory have been made as a result of the intensity interferometry problem, and we refer the reader to Mandel⁶ for a discussion of this subject.

The quantum aspects of coherence theory have also been brought out by the need to consider fourth-order moments when studying laser radiation. The radiation from stars was generally accepted to have gaussian statistics and thus a simple relationship between $I'(P_1, P_2, 0)$ and the scalar analog of $R_{ik}(P_1, P_2, 0)$ is expected. The statistical character of laser light is presently unknown however, and the fourth-order moment is expected to shed new light on the character of laser radiation rather than providing a more convenient measurement as it sometimes did in star measurements. A quantum field formalism for coherence theory intended to consider all types of radiation has been studied by a number of authors (Glauber, Sudarshan, Wolf), and a series of pertinent papers will be found in the proceedings of the Quantum Electronic Conference.⁷

Applications (Measurement of the Angular Diameter of Stars). Coherence theory is of great use in the treatment of imaging and mapping problems. It is especially convenient for treating the problems involving the effects of the atmosphere on resolution. To understand phenomena like the twinkle of stars, it is a natural formalism to introduce. To present the reader with a definite example of the use of the coherence theory formalism, we will conclude this brief discussion with an outline of the measurement of the angular diameter of visible stars.

Let us suppose that the source in Fig. 1 is a visible star so that l_1 is many light-years. There is no telescope big enough to resolve any star, and to make direct measurements of the angular diameter of a star, Michelson introduced the use of interference experiments to essentially give one a bigger effective aperture.

For purposes of visible star measurement we can replace the star of diameter D by a circular disk of diameter D lying in a plane parallel to A . The radiation leaving the star is assumed to be incoherent so that for all points on the disk we may take $I'(P_1, P_2, 0) = c\delta(P_1 - P_2)$. Using Eq. (4) this allows us to solve for $I'(P_1, P_2, 0)$ on the earth (screen A) if we filter the starlight to insure the validity of the quasi-monochromatic

approximation. We find

$$\Gamma(P_1, P_2, 0) = \text{const.} \frac{J_1 \left[\frac{kdD/2}{l_1} \right]}{\frac{kdD/2}{l_1}} \quad (10)$$

where k is the average wave number of the light and J_1 is a first-order Bessel function.

If we define the angular diameter of the star as $\theta = D/l_1$, we see that J_1 equals zero when $\theta = 1.22\lambda/d$, where $\lambda = 2\pi/k$. When J_1 equals zero, there are no fringes on the screen B. Hence to find the angular diameter of the star, we need only increase d from zero, when there will be high contrast fringes, to the separation d when there are no fringes. Putting this latter value into our expression for θ gives us the angular diameter of the star.

Basic References. For a fundamental treatment of coherence theory we refer the reader to the following basic texts: Born and Wolf,² O'Neill,⁸ and Beran and Parrent.¹

MARK J. BERAN

References

1. Beran, M., and Parrent, G., Jr., "Theory of Partial Coherence," Englewood Cliffs, N. J., Prentice Hall, 1964.
2. Born, M., and Wolf, E., "Principles of Optics," London, Pergamon Press, 1959.
3. Grivet, P., and Bloembergen, N., Eds., "Quantum Electronics, Proceedings, of Third International Congress, Paris," New York, Columbia University Press, 1964.
4. Hanbury Brown, R., and Twiss, R., *Proc. Royal Soc. London Ser. A*, **242**, 300 (1957).
5. Hanbury Brown, R., and Twiss, R., *Proc. Royal Soc. London Ser. A*, **243**, 291 (1957).
6. Mandel, L., "Progress in Optics," Vol. 3, p. 181, Amsterdam, North Holland Publishing Co., 1963.
7. Michelson, A., *Phil. Mag.* (5), **30**, 1 (1890).
8. O'Neill, E., "Introduction to Statistical Optics," Reading, Mass., Addison-Wesley, 1963.
9. Wolf, E., *Proc. Royal Soc. London Ser. A*, **230**, 246 (1955).

Cross-references: INTERFERENCE AND INTERFEROMETRY, LASER, MASER.

COLLISIONS OF PARTICLES

Introduction. Scattering experiments provide the principal technique by which physicists attempt to understand the structure and interactions of matter on a microscopic scale. Scattering theory provides the basis for analyzing and interpreting scattering experiments.

A description of the development of scattering theory may be divided into several topics. The oldest and simplest branch of scattering theory is that of potential scattering, or scattering of two particles which interact through a *local potential*¹. Potential scattering was studied extensively in the first two decades following the development of

quantum mechanics in the analysis of elastic scattering of particles by atoms and of nucleon-nucleon scattering. The latter topic, in particular, led to the introduction of an elaborate theory of scattering by noncentral interactions². The development of nuclear physics, with the observation of resonance reactions, indicated the need for more general descriptions of scattering. The resulting theory of resonance reactions^{3,4} has leaned only rather lightly on the details of the Schrödinger equation. Quantum field theory was developed to describe electromagnetic phenomena. Of major importance in the development of scattering theory was the introduction of renormalization techniques into FIELD THEORY.⁶

It might be claimed that modern scattering theory began with the integral equation formulation of Lippmann and Schwinger and the introduction of S-matrix theory by Heisenberg⁶ and others⁹. This work has stimulated much of the development of theoretical physics in the last decade. Of particular significance are the clarification of the study of rearrangement collisions and the development of the so-called dispersion theoretic techniques.

The Scattering Cross Section. The properties of scattering interactions are usually expressed most conveniently in terms of the *scattering cross section*. To define this term, we consider the following scattering experiment: A beam of particles (called *beam particles*) is directed on a scatterer consisting of *target particles*. As a result of collisions between beam and target particles, there are particles which emerge from the reactions (called *reaction products*) and these are detected in *particle detectors*. To describe this quantitatively, we suppose that the scatterer contains N_t target particles and that this is uniformly illuminated by a flux F_{11} (expressed as the number of beam particles per unit area per unit time arriving at the target) of beam particles. We suppose also that the scatterer is sufficiently small that the beam is negligibly attenuated in passing through it. Then, if there are δN_t scattering interactions per unit time which lead to detected particles, we define the scattering cross section $\delta\sigma$ as:¹⁰

$$\delta\sigma = \frac{\delta N_t}{N_t F_{11}} \quad (1)$$

In the limit that the detectors subtend very small solid angles, as seen from the target, we define the *differential scattering cross section* $d\sigma$. When a single detector, subtending a solid angle $\delta\Omega$, is used to define $\delta\sigma$, we may define the cross section per unit solid angle as

$$\frac{d\sigma}{d\Omega} = \lim_{\delta\Omega \rightarrow 0} \frac{\delta\sigma}{\delta\Omega} \quad (2)$$

The total scattering cross section σ is obtained by summing $\delta\sigma$ over all scattering events:

$$\sigma = \sum \delta\sigma \quad (3)$$

[Equation (3) does not exist for scattering by a coulomb force.]

The scattering cross section may be expressed in terms of the square of the magnitude of a scattering amplitude (or S-matrix, or T-matrix element) and is completely described as a function of the momenta and internal states of the particles in the initial and final states. Thus, for the two particles prior to collision¹¹ we may take the momenta \mathbf{p}_1 and \mathbf{p}_2 and the internal state quantum numbers s_1 and s_2 as variables. (For example, s_1 and s_2 may describe spin orientation, isotopic spin, etc. For colliding molecules these variables will describe vibrational, rotational and electronic states.) We may suppose there to be μ particles in the final state following the collision and specify this state by the momenta and internal variables $\mathbf{k}_1, \dots, \mathbf{k}_\mu, s_1', \dots, s_\mu'$. The scattering cross section may be expressed in terms of these variables. Because of symmetries, the number of variables required to describe $\delta\sigma$ may ordinarily be reduced. The most commonly encountered of these symmetries are: (1) energy and momentum conservation; (2) rotational invariance; (3) the Lorentz invariance of the scattering cross section $\delta\sigma$.¹² The Lorentz invariance of $\delta\sigma$ permits one to describe the scattering in the *barycentric* coordinate system—the coordinate system in which the total momentum of the interacting particles is zero.

This may be illustrated for the special case for which there are only two particles in the initial and two in the final state and an average has been performed over all spin orientations. Then the twelve components of $\mathbf{p}_1, \mathbf{p}_2, \mathbf{k}_1$ and \mathbf{k}_2 may be replaced by only two variables. These may be taken, for example, as the barycentric energy and angle between \mathbf{p}_1 and \mathbf{k}_1 . Convenient variables in relativistic analyses are often chosen to be the Lorentz invariants

$$s \equiv (p_1 + p_2)^2 = (k_1 + k_2)^2$$

$$t \equiv (p_1 - k_1)^2 = (p_2 - k_2)^2 \quad (4)$$

where we have written p_1 , etc., for the four-component energy-momentum vector.

Potential Scattering. We briefly illustrate the discussion of the preceding section with the example of nonrelativistic scattering by a local central potential $V(r)$. The SCHRÖDINGER EQUATION for scattering in the barycentric coordinate system is

$$[\nabla^2 + \kappa^2 - v(r)]\psi_{\mathbf{k}}(\mathbf{r}) = 0 \quad (5)$$

Here $\hbar\kappa$ is the momentum of particle "1" in the barycentric system and $v(r) = \frac{2M_r}{\hbar^2} V(r)$, with M_r the reduced mass of the two particles. In the limit of large separation r between the particles the wavefunction $\psi_{\mathbf{k}}$ has the asymptotic form [our notation is such that we represent a unit vector in the direction of \mathbf{k} by $\hat{\mathbf{k}}$]

$$\psi_{\mathbf{k}}(\mathbf{r}) \rightarrow (2\pi)^{-3/2} \left[e^{i\mathbf{k}\cdot\mathbf{r}} + \frac{e^{i\kappa r}}{r} f(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) \right] \quad (6)$$

Here $f(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})$ is the scattering amplitude for scattering particle "1" from the direction $\hat{\mathbf{k}}$ into the direction $\hat{\mathbf{r}}$. The corresponding cross section per unit solid angle is

$$\frac{d\sigma}{d\Omega} = |f(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}})|^2 \quad (7)$$

The wave function $\psi_{\mathbf{k}}$ may be expanded into partial waves as follows:

$$\psi_{\mathbf{k}}(\mathbf{r}) = \sum_{l=0}^{\infty} \frac{(2l+1)}{4\pi\kappa r} P_l(\hat{\mathbf{k}}\cdot\hat{\mathbf{r}}) i^l e^{i\delta_l} w_l(\kappa; r) \quad (8)$$

Here P_l is the Legendre polynomial of order l , δ_l is the scattering phase shift [see Eq. (10) below], and $w_l(\kappa; r)$ satisfies the differential equation [Eq. (13)]

$$\left[\frac{d^2}{dr^2} + \kappa^2 - \frac{l(l+1)}{r^2} - v(r) \right] w_l = 0 \quad (9)$$

This is to be integrated subject to the condition that w_l is regular at $r = 0$. For large r , w_l has the asymptotic form

$$w_l(\kappa; r) \sim \sqrt{\frac{2}{\pi}} \sin\left(\kappa r - \frac{\pi l}{2} + \delta_l\right) \quad (10)$$

It is Eq. (10) which permits the determination of the phase shift δ_l . The quantity

$$S_l(\kappa) = \exp[2i\delta_l(\kappa)] \quad (11)$$

is an eigenvalue of Heisenberg's S-matrix.⁸

For scattering by noncentral forces, the potential $V(\mathbf{r}, S_1, S_2)$ is a function of \mathbf{r} (and sometimes the orbital angular momentum operator) and the spin operators S_1 and S_2 of the two colliding particles (if either has no spin, we consider its spin operator to be zero). Spin eigenfunctions $u(\nu_1, \nu_2)$ may be introduced as depending on the orientations ν_1 and ν_2 of the respective spins of magnitudes S_1 and S_2 . Then the wave function $\psi_{\mathbf{k}, \nu_1, \nu_2}$ is to be labeled with the initial spin orientations ν_1 and ν_2 . The asymptotic form corresponding to Eq. (6) is

$$\psi_{\mathbf{k}, \nu_1, \nu_2} \rightarrow (2\pi)^{-3/2} \left[e^{i\mathbf{k}\cdot\mathbf{r}} u(\nu_1, \nu_2) + \frac{e^{i\kappa r}}{r} \sum_{\nu_1', \nu_2'} \langle \nu_1', \nu_2' | f(\hat{\mathbf{k}}, \hat{\mathbf{r}}) | \nu_1, \nu_2 \rangle u(\nu_1', \nu_2') \right] \quad (12)$$

Here $\langle \nu_1', \nu_2' | f(\hat{\mathbf{k}}, \hat{\mathbf{r}}) | \nu_1, \nu_2 \rangle$ is the scattering amplitude for scattering to a final spin orientation ν_1', ν_2' . The cross section per unit solid angle is in this case

$$\frac{d\sigma}{d\Omega} = |\langle \nu_1', \nu_2' | f(\hat{\mathbf{k}}, \hat{\mathbf{r}}) | \nu_1, \nu_2 \rangle|^2 \quad (13)$$

For an unpolarized initial state, corresponding to a uniform mixture of the $(2S_1 + 1)(2S_2 + 1)$ spin states, the cross section for scattering particle

"1" into the direction $\hat{\mathbf{r}}$ with any spin orientation is

$$\frac{d\bar{\sigma}}{d\Omega} = \frac{1}{(2S_1 + 1)(2S_2 + 1)} \sum_{\nu_1', \nu_2'} \sum_{\nu_1, \nu_2} |\langle \nu_1', \nu_2' | f | \nu_1, \nu_2 \rangle|^2 \quad (14)$$

where the sums extend over all spin orientations.

Following scattering by noncentral forces, the particles will in general have preferred spin orientations, or be *polarized*. When, for example, particle "1" has spin one-half with a spin operator σ_1 , we define its polarization vector $\mathbf{P}(\nu_1, \nu_2)$ by the equation

$$\mathbf{P}(\nu_1, \nu_2) = \left\{ \sum_{\nu_1'', \nu_2'} [\langle \nu_1'', \nu_2' | f | \nu_1, \nu_2 \rangle]^* \langle \nu_1'' | \sigma_1 | \nu_1' \rangle \langle \nu_1', \nu_2' | f | \nu_1, \nu_2 \rangle \right\} \times \left\{ \sum_{\nu_1', \nu_2'} |\langle \nu_1', \nu_2' | f | \nu_1, \nu_2 \rangle|^2 \right\}^{-1} \quad (15)$$

For an unpolarized initial state, the polarization is

$$\bar{\mathbf{P}} = \frac{1}{(2S_1 + 1)(2S_2 + 1)} \sum_{\nu_1, \nu_2} \mathbf{P}(\nu_1, \nu_2) \quad (16)$$

The study of polarization following scattering has provided an important tool for analyzing nuclear and elementary particle reactions.^{14,15} In particular, the role of noncentral interactions in nucleon-nucleon scattering has been studied in great detail.¹⁶

Formal Scattering Theory. To describe a general scattering reaction Lippmann and Schwinger^{7,17} introduced a scattering matrix \mathcal{T}_{ba} to describe scattering from an initial state χ_a to a final state χ_b .¹⁸ This is defined as

$$\mathcal{T}_{ba} = (X_b, V\psi_a) \quad (17)$$

where ψ_a is the steady-state wave function for the event and V is the scattering interaction. Since momentum is conserved for an isolated scattering, we may write

$$\mathcal{T}_{ba} = \delta(\mathbf{P}_b - \mathbf{P}_a) T_{ba} \quad (18)$$

where \mathbf{P}_a and \mathbf{P}_b are the total momenta of the particles in the initial and final states, respectively, and T_{ba} is defined only for states b and a corresponding to $\mathbf{P}_b = \mathbf{P}_a$.

The scattering cross section $d\sigma$ [Eq. (1)] is expressed in terms of T_{ba} as¹²

$$d\sigma = \frac{(2\pi)^4}{v_{\text{rel}}} \sum_b \delta(\mathbf{P}_b - \mathbf{P}_a) \delta(E_b - E_a) |T_{ba}|^2 \quad (19)$$

Here v_{rel} is the relative velocity of beam and target particles, E_b and E_a are the respective total energies of the particles in states b and a , and the sum on b extends over those states which lead to the reaction products striking the detectors and

thus to register an event. We emphasize that Eq. (19) is Lorentz invariant.¹²

The Heisenberg S-matrix⁸ is given by the expression

$$S_{ba} = \delta_{ba} - 2\pi i \delta(E_b - E_a) \mathcal{T}_{ba} \quad (20)$$

where δ_{ba} is a Dirac δ -function. The S-matrix is unitary, so

$$\sum_c S_{cb}^\dagger S_{ba} = \delta_{ca} \quad (21)$$

On substituting Eq. (20) into this, we obtain the equivalent expression of unitarity

$$i[\mathcal{T}_{ca} - \mathcal{T}_{cb}^\dagger] = 2\pi \sum_b \mathcal{T}_{cb}^\dagger \delta(E_b - E_a) \mathcal{T}_{ba} \quad (22)$$

which is defined only for states c and a on the same *energy shell* (corresponding to $E_c = E_a$).

The fundamental problem of scattering theory is to determine the \mathcal{T} -matrix on the energy shell (or, equivalently, the S-matrix). The first step in doing this is to make use of general symmetry principles (such as Lorentz invariance) to limit the functional forms allowed. Following this a dynamical principle is needed. Such dynamical principles (reviewed in Chapters 5 and 10 of reference¹⁰) have been proposed in a great variety of forms including integral equations, variational principles, and conditions of functional analyticity.

Rearrangement collisions (i.e., collisions in which bound particles rearrange themselves) have been studied extensively following the development of formal scattering theory. Much of this¹⁹ was stimulated by the observance of apparent paradoxes²⁰. An interesting modification of Eq. (17) intended for application to rearrangement collisions has been given by Mittleman.²¹

Another, and not entirely unrelated, class of applications of formal scattering theory concerns scattering by composite systems. These include the multiple scattering and optical model descriptions²² and elaborate theories of atomic scattering processes^{23,24}. The successful development and use of variational principles for such processes should also be noted.²⁵

Field Theory. Quantum field theory was originally developed to describe electro-magnetic phenomena. It was applied in a promising context during the 1930's to β -decay and to the meson theory of nuclear forces. The great optimism following the development of renormalization theory⁶ faded quickly for want of adequate mathematical techniques for handling strong interactions. The most successful applications to strong interactions were the semi-phenomenological calculations of Chew and others^{5,26}.

An interesting and novel attempt to revive field theory has been initiated by Weinberg.²⁷

S-matrix Theory. Heisenberg suggested in 1946⁸ that a proper quantum theory of scattering would deal only with observable quantities such as the S-matrix and should not require off-the-energy-shell matrix elements of such quantities as \mathcal{T} [Eq. (18)]. Considerable impetus for this point

of view has been given by the development of *dispersion theory*, following early suggestions of Wigner and others.²⁸ The first attempt at a systematic formulation of a dispersion relation within the context of quantum field theory was made by Gell-Mann, Goldberger, and Thirring.²⁹ Further development followed applications of formal scattering theory to quantum field theory.³⁰ The development of the Mandelstam representation³¹ provided an important step toward obtaining a "dynamical principle." A further important step was the proposal by Chew and Frautschi and Blankenbecker and Goldberger,³² who suggested that the only singularities of the S-matrix are those required by the unitarity condition [Eq. (22)] and that families of particles should be associated with Regge Trajectories.³⁴

KENNETH M. WATSON

References

1. The early development of scattering theory is well described in the classic work of Mott, N. F., and Massey, H. S. W., "The Theory of Atomic Collisions," Oxford, Clarendon Press, 1933.
2. Rarita, W., and Schwinger, J., *Phys. Rev.*, **59**, 436 (1941). Christian, R. S., and Hart, E. W., *Phys. Rev.*, **77**, 441 (1950). Christian, R. S., and Noyes, H. P., *Phys. Rev.*, **79**, 85 (1950).
3. Breit, G., and Wigner, E. P., *Phys. Rev.*, **49**, 519, 642 (1936).
4. Wigner, E. P., *Phys. Rev.*, **70**, 15, 606 (1946); Wigner, E. P., and Eisenbud, L., *Phys. Rev.*, **72**, 29 (1947).
Sachs, R. G., "Nuclear Physics," Reading, Mass., Addison-Wesley Publishing Co., 1953.
5. See, for example, Mandl, F., "Introduction to Quantum Field Theory," New York, Interscience Publishers, 1959. The older work is admirably described in G. Wentzel, "Quantum Theory of Fields," New York, Interscience Publishers, 1949.
6. Feynman, R. P., *Phys. Rev.*, **76**, 749 (1949). Dyson, F. J., *Phys. Rev.*, **75**, 486 (1949). Tomonaga, S., *Progr. Theoret. Phys. (Kyoto)*, **1**, 27 (1946). Schwinger, J., *Phys. Rev.*, **74**, 1439 (1948).
7. Lippmann, B., and Schwinger, J., *Phys. Rev.*, **79**, 469 (1950).
8. Heisenberg, W., *Z. Naturforsch.*, **1**, 608 (1946).
9. Wheeler, J. A., *Phys. Rev.*, **52**, 1107 (1937). Moller, C., *Kgl. Danske Videnskab. Selskab, Mat. Fys. Medd.*, **23**, 1 (1948).
10. This is a much abbreviated version of the discussion given in Ch. 3 of Goldberger, M. L., and Watson, K. M., "Collision Theory," New York, John Wiley & Sons, Inc., 1964.
11. The case that more than two particles collide is important for the discussion of chemical reactions in gases and liquids. This is discussed, for example, in Ch. 5 and Appendix B of reference 10.
12. See, for example, p. 90 of reference 10.
13. Following the notation of Section 6.3, reference 10.
14. Wolfenstein, L., and Ashkin, J., *Phys. Rev.*, **85**, 947 (1952). Simon, A., and Welton, T., *Phys. Rev.*, **93**, 1435 (1954). Wolfenstein, L., *Ann. Rev. Nucl. Sci.*, **6**, 43 (1956).
15. A comprehensive account of the theory is given in Ch. 7 of reference 10.
16. Moravcsik, M. J., and Noyes, H. P., "Theories of Nucleon-Nucleon Elastic Scattering," *Ann. Rev. Nucl. Sci.*, **11**, 95 (1961).
17. Gell-Mann, M., and Goldberger, M. L., *Phys. Rev.*, **91**, 398 (1953).
18. We are here following the notation of Chs. 3 and 5 of reference 10.
19. Brening, W., and Haag, R., *Fortschr. Physik*, **7**, 183 (1959). Ekstein, H., *Phys. Rev.*, **101**, 880 (1956). Cook, J., *J. Math. Phys.*, **36**, 82 (1957).
A somewhat more flexible interpretation has been made in Ch. 4 of reference 10.
20. Foldy, L., and Tobocon, W., *Phys. Rev.*, **105**, 1099 (1957).
Eptein, S., *Phys. Rev.*, **106**, 598 (1957).
Lippmann, B., *Phys. Rev.*, **102**, 264 (1956).
21. Mittelman, M. H., *Phys. Rev.*, **122**, 1930 (1961); **126**, 373 (1962).
22. Watson, K. M., *Phys. Rev.*, **89**, 575 (1953); **105**, 1388 (1957).
Francis, N. C., and Watson, K. M., *Phys. Rev.*, **92**, 291 (1953). See also, Ch. 11 of reference 10.
23. Massey, H. W. W., "Encyclopedia of Physics," Berlin-Göttinger-Heidelberg, Springer-Verlag, 1956.
24. Burke, P. G., and Smith, K., *Rev. Mod. Phys.*, **34**, 458 (1962).
Malik, F. B., and Treffitz, E. T., *Z. Astrophys.*, **50**, 96 (1960).
Lippmann, B., Mittleman, M., and Watson, K., *Phys. Rev.*, **116**, 920 (1959).
Temkin, A., and Pohle, R., *Phys. Rev. Letters*, **10**, 268 (1963) and earlier references.
25. Schwartz, C., *Phys. Rev.*, **124**, 1468 (1961).
O'Malley, T. F., Sprach, L., and Rosenberg, L., *J. Math. Phys.*, **2**, 491 (1961) and earlier references.
Sugar, R., and Blankenbecker, R., *Phys. Rev.*, **136**, B472 (1964).
26. This subject is reviewed by Wick, G. C., *Rev. Mod. Phys.*, **27**, 339 (1955).
27. Weinberg, S., *Phys. Rev.*, **130**, 776 (1963).
28. The history of this subject is reviewed in Ch. 10 of reference 10.
29. Gell-Mann, M., Goldberger, M. L., and Thirring, W., *Phys. Rev.*, **95**, 1612 (1954).
30. Lehmann, H., Symanzik, K., and Zimmerman, W., *Nuovo Cimento*, **1**, 205 (1955).
31. Mandelstam, S., *Phys. Rev.*, **112**, 1344 (1955); **115**, 1741, 1759 (1959).
32. Chew, G. F., and Frautschi, S. C., *Phys. Rev. Letters*, **8**, 41 (1962).
33. Blankenbecker, R., and Goldberger, M. L., *Phys. Rev.*, **126**, 766 (1962).
34. Reviews of the S-matrix theory of scattering can be found in Chew, G. F., "S-Matrix Theory of Strong Interactions," New York, Benjamin, 1961.
Omnes, R., and Froissart, M., "Mandelstam Theory and Regge Poles," New York, Benjamin, 1963.

Cross-references: ATOMIC AND MOLECULAR BEAMS, CONSERVATION LAWS AND SYMMETRY, CROSS SECTIONS AND STOPPING POWER, FIELD THEORY, SCHRÖDINGER EQUATION.

COLOR

Definition. Color is the property of light by which an observer may discriminate between two structure-free patches of light of identical size and shape. If two such patches cannot be distinguished by eye from each other, they are said to match in color.

Conditions for a Color Match. Two patches of light defined physically by their spectral concentrations ($dL/d\lambda$) of radiance L (radiant energy per unit time, unit solid angle, and unit of orthogonally projected area) color match provided that simultaneously:

$$\begin{aligned}\int \bar{x}_\lambda(dL/d\lambda)_1 d\lambda &= \int \bar{x}_\lambda(dL/d\lambda)_2 d\lambda \\ \int \bar{y}_\lambda(dL/d\lambda)_1 d\lambda &= \int \bar{y}_\lambda(dL/d\lambda)_2 d\lambda \\ \int \bar{z}_\lambda(dL/d\lambda)_1 d\lambda &= \int \bar{z}_\lambda(dL/d\lambda)_2 d\lambda\end{aligned}\quad (1)$$

where \bar{x}_λ , \bar{y}_λ , \bar{z}_λ are the color-matching functions of wavelength λ characterizing the observer. Values of these functions customarily used for observers of average normal color vision are those recommended in 1931 by the International Commission of Illumination (CIE); with trivial exceptions any weighted mean of \bar{x}_λ , \bar{y}_λ , \bar{z}_λ , may be substituted for any of them without changing the meaning of Eq. (1). If the two light patches have identical spectral concentrations of radiance, Eq. (1) is satisfied for all observers and describes a nonmetameric match. If Eq. (1) is satisfied even though $(dL/d\lambda)_1$ is different from $(dL/d\lambda)_2$ for some parts of the visible spectrum, Eq. (1) describes a metameric match. Such a color match will usually be a mismatch for observers not characterized by \bar{x}_λ , \bar{y}_λ , \bar{z}_λ .

Deviations from normal color vision are of three types depending on the number of requirements for a color match that have to be satisfied. Anomalous trichromatic vision (protanomalous vision, deuteranomalous vision), like normal trichromatic vision, requires satisfaction of three conditions. Dichromatic vision (protanopia,

deuteranopia, and tritanopia) requires satisfaction of but two conditions, and monochromatic vision (characterizing total color blindness and also characterizing normal vision in sufficiently dim light) requires satisfaction of but one condition. The color-matching function for the usual type of monochromatic vision is the CIE scotopic luminous-efficiency function recommended in 1951. The color-matching functions for the other forms of deviant color vision are related to those (\bar{x}_λ , \bar{y}_λ , \bar{z}_λ) for normal daylight vision as given in Table 1. To make clear this relation, the color-matching functions for normal vision are stated, as is admissible, in terms of $-.460\bar{x}_\lambda + 1.359\bar{y}_\lambda + 0.101\bar{z}_\lambda$ instead of \bar{x}_λ . An alternate statement, to be discussed later, is also given. Figure 1 shows all of these color-matching functions.

Measurement of Color. The integrals of Eq. (1), called tristimulus values, X , Y , Z , serve as specifications of the color of the light patch defined physically by the values of spectral concentration of radiance ($dL/d\lambda$) throughout the visible spectrum. The tristimulus values may be determined visually by having the observer adjust a mixture of three lights, called working primaries, to produce a color match for the specimen light patch. The required amounts of the working primaries, R , G , B , are tristimulus values relative to the working primaries. The tristimulus values, X , Y , Z , relative to the CIE primaries may then be computed from the tristimulus values (X_r , Y_r , Z_r ; X_g , Y_g , Z_g ; X_b , Y_b , Z_b) of the working primaries, thus:

$$\begin{aligned}X &= X_r R + X_g G + X_b B \\ Y &= Y_r R + Y_g G + Y_b B \\ Z &= Z_r R + Z_g G + Z_b B\end{aligned}\quad (2)$$

Photoelectric color measurement may be accomplished by providing a photocell with three filters such that its spectral sensitivity may be made proportional in turn to \bar{x}_λ , \bar{y}_λ , \bar{z}_λ . When the light patch of color to be determined is projected

TABLE 1. COLOR-MATCHING FUNCTIONS FOR VARIOUS TYPES OF TRICHROMATIC AND DICHROMATIC VISION

Type of Vision	Color-matching Function		
	Long-wave Sensitive	Middle-wave Sensitive	Short-wave Sensitive
Trichromatic			
Normal (1, 5, 7)*	$(.639\bar{x}_\lambda + .490\bar{y}_\lambda - .129\bar{z}_\lambda)$	$(-.509\bar{x}_\lambda + 1.410\bar{y}_\lambda + .099\bar{z}_\lambda)$	\bar{z}_λ
(4, 5, 7)*	\bar{y}_λ	$(.460\bar{x}_\lambda + 1.359\bar{y}_\lambda + .101\bar{z}_\lambda)$	\bar{z}_λ
Protanomalous (2, 5, 7)*	$(.32\bar{x}_\lambda + .25\bar{y}_\lambda - .07\bar{z}_\lambda)n^{**}$	$(-.460\bar{x}_\lambda + 1.359\bar{y}_\lambda + .101\bar{z}_\lambda)$	\bar{z}_λ
Deuteranomalous (2, 3-4, 7)*	$(.32\bar{x}_\lambda + .25\bar{y}_\lambda - .07\bar{z}_\lambda)n^{**}$	Intermediate to \bar{y}_λ and $(.312\bar{x}_\lambda + .757\bar{y}_\lambda - .069\bar{z}_\lambda)$	\bar{z}_λ
Dichromatic			
Protanopic (5, 7)*	None	$(-.460\bar{x}_\lambda + 1.359\bar{y}_\lambda + .101\bar{z}_\lambda)$	\bar{z}_λ
Deuteranopic (3-4, 7)*	Intermediate to \bar{y}_λ and $(.312\bar{x}_\lambda + .757\bar{y}_\lambda - .069\bar{z}_\lambda)$	None	\bar{z}_λ
Tritanopic (4, 5)*	\bar{y}_λ	$(-.460\bar{x}_\lambda + 1.359\bar{y}_\lambda + .101\bar{z}_\lambda)$	None

* Numbers of the curves of Fig. 1 showing these color-matching functions.

** The number, n , greater than one, characterizes each individual anomalous trichromatic observer.

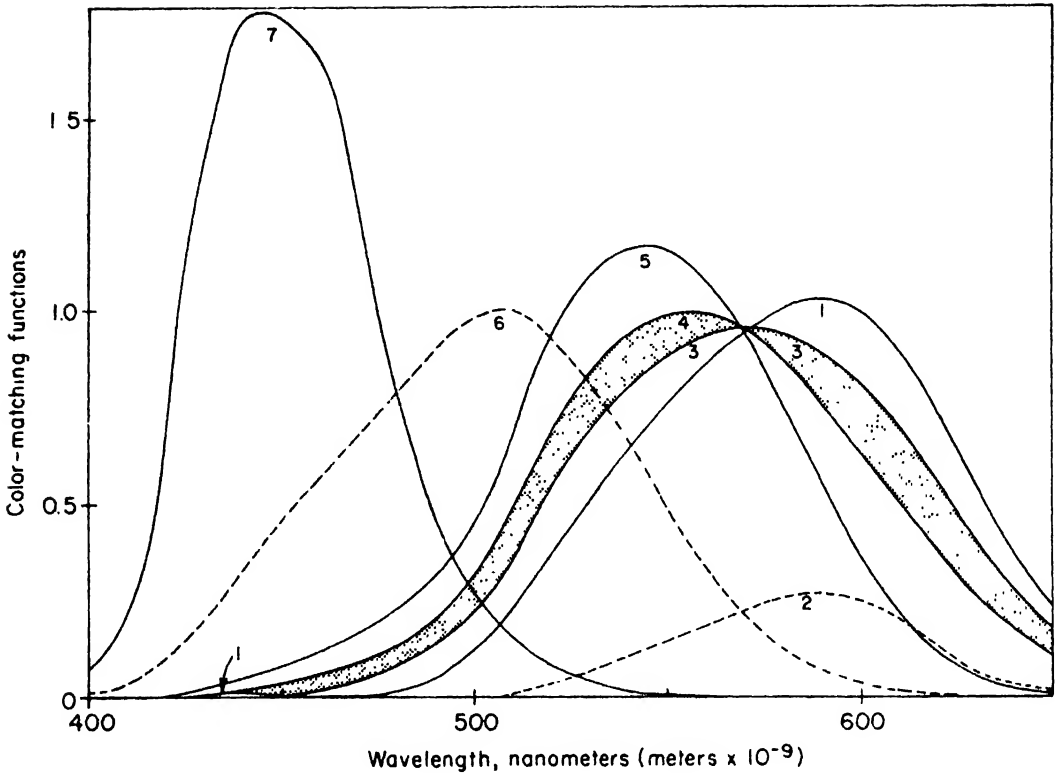


Fig. 1. Color-matching functions for normal and various forms of deviant vision:
 Curve 1 = $(.639\bar{x}_\lambda + .490\bar{y}_\lambda - .029\bar{z}_\lambda)$, product of spectral transmittance of eye media by spectral absorptance of photo-pigment presumed to exist in the retina.
 Curve 2 = $(.32\bar{x}_\lambda + .25\bar{y}_\lambda - .07\bar{z}_\lambda)^2$, long-wave-sensitive function for both protanomalous and deuteranomalous vision ($n=2$).
 Curve 3 = $(.312\bar{x}_\lambda + .757\bar{y}_\lambda - .069\bar{z}_\lambda)$, long-wave extreme of the range (shown crosshatched) or deuteranopic long-wave functions; also the long-wave extreme of the deuteranomalous middle-wave functions.
 Curve 4 = \bar{y}_λ , photopic luminous-efficiency function, which is the long-wave-sensitive function for normal and tritanopic vision, and the short-wave extreme of the range (shown crosshatched) of long-wave-sensitive deuteranopic functions and of the middle-wave deuteranomalous functions.
 Curve 5 = $(-.460\bar{x}_\lambda + 1.359\bar{y}_\lambda + .101\bar{z}_\lambda)$, long-wave-sensitive protanopic function, middle-wave-sensitive function for protanomalous and normal color vision, and short-wave-sensitive function for tritanopic vision; also close to $(-.509\bar{x}_\lambda + 1.410\bar{y}_\lambda + .099\bar{z}_\lambda)$, the product of spectral transmittance of eye media by spectral absorptance of photo-pigment presumed to exist in the retina.
 Curve 6 = scotopic luminous-efficiency function, the color-matching function for the usual type of total color blindness.
 Curve 7 = \bar{z}_λ , short-wave-sensitive function for all types of vision except tritanopic and scotopic; also product of the spectral transmittance of the eye media by the spectral absorptance of photo-pigment presumed to exist in the retina.

onto the photosensitive surface with the X -tristimulus filter interposed, the response of the cell will be proportional to $\int (dL/d\lambda)\bar{x}d\lambda = X$; and by inserting the Y -tristimulus filter and the Z -tristimulus filter in turn, Y and Z may be found for the unknown color.

Color Perception and Chromatic Adaptation. The color perception of a spot of light against a dark surround may be described in terms of hue, brightness and saturation. Hue refers to red, orange, yellow, green, blue, purple, back to red, and their intermediates. Brightness determines

whether the patch appears to be emitting more or less light per unit area; it varies from very dim to dazzling. Saturation refers to the amount of difference between the color perception to be described and the neutral color perception (exhibiting no hue) of the same brightness; it varies from neutral to very vivid. Average daylight usually yields a neutral color perception. The color perception of an object illuminated by daylight may be described in terms of hue, saturation (see above) and lightness. Lightness determines whether the object appears to be reflecting a

greater or lesser fraction of the incident light; it varies from a minimum for black to a maximum for white.

The color perception of a spot of light viewed against a dark background correlates well with the tristimulus values of the spot. For example, brightness corresponds to the tristimulus value Y , also known as luminance, but with nonlinear scaling. The cube-root of luminance $Y^{1/3}$ correlates well with estimates of brightness.

The color perception of an object illuminated by average daylight and viewed by a daylight-adapted observer against a white background correlates well with the tristimulus values of the light reflected toward the observer's eye from the object relative to that from the surround. For example, lightness corresponds to Y/Y_s , but with nonlinear scaling where Y_s refers to the surround. The function $(Y/Y_s)^{1/3}$ correlates well with estimates of lightness.

If the light patch or object is viewed by an observer adapted to a color quite different from darkness or daylight (such as by use of a vivid red surround) the color perceptions are markedly changed. If the tristimulus values of the color which appears gray to the observer in his changed state of adaptation be X_a, Y_a, Z_a , and that of the color which appears gray in daylight adaptation be X_1, Y_1, Z_1 , the ratios $X_1/X_a, Y_1/Y_a$, and Z_1/Z_a serve as measures of the changed state of adaptation. The color perceptions then correlate to a first degree of approximation to $(X_1/X_a)X$, $(Y_1/Y_a)Y$, and $(Z_1/Z_a)Z$, instead of to X, Y, Z . Still better approximations are obtained if primaries different from the CIE primaries are used, such as those implied by curves 3, 5, and 7 of Fig. 1. The approximate correlation so achieved is known as the v. Kries coefficient law.

Control of Color by Pigments and Dyes. Objects owe their color to materials (colorants) incorporated in them or spread over their surfaces. Objects viewed by transmitted light usually owe their colors to soluble colorants (dyes). The spectral concentration $dE/d\lambda$, of the radiant flux incident per unit area is changed to $T_\lambda(dE/d\lambda)$ on passage through the dyed object, where T_λ is the spectral transmittance of the object. Objects viewed by reflected light usually owe their colors to insoluble colorants (pigments) which scatter as well as transmit the incident light. The light penetrates the object, and after multiple scattering emerges from the illuminated side. The spectral concentration, $dE/d\lambda$, is changed to $R_\lambda(dE/d\lambda)$, where R_λ is the spectral reflectance of the object.

The colors of objects may be specified by the tristimulus values of the light leaving them toward the observer's eye compared to the tristimulus values of the incident light.

Color Scales. A series of color standards whose colors are suitably spaced over a color range form a color scale. Perhaps the most common example of a one-dimensional color scale is a series of solutions of different colors produced by different concentrations of a known colorant. By visual comparison of a solution of unknown concentration of this colorant with the color standards

making up the scale, the concentration of the unknown may be evaluated. Another one-dimensional color scale is a series of grays ranging from black to white, and still another is the scale of color temperature, defined as the temperature of a blackbody required to make it match the color of an unknown light source to be evaluated.

If the color range is two dimensional, a two-dimensional color scale consisting of a family of one-dimensional scales is required. An example is the range of colors producible by mixing in various proportions paints of any three colors, say black, white, and red. The colors produced by mixing black paint with white paint in various proportions form one of the series of one-dimensional scales, and the other one-dimensional scales might be produced with 10 per cent red paint, 20 per cent, and so on.

If the color range is three dimensional, a three-dimensional color scale is required. The color standards of the Lovibond, the Ostwald, and the Munsell color systems are examples of three-dimensional color scales; the Army solutions form another.

Automatic Production of Pictures in Color. Pictures produced automatically in color as in television, photography, or multiple printing, depend on the principle of photoelectric color measurement. A record of the tristimulus values of each picture element is made either by the television camera or by a camera with photographic film. The final picture is synthesized by the information stored in this record. The synthesis is made by averaging the colors of small spots (television), by addition of colorant layers through which light passes in succession on its way to the observer's eye (most color photography), or by a combination of the two (process printing). The tristimulus values X_a, Y_a, Z_a , of the average color of a group of juxtaposed spots too small to be resolved by eye can be computed simply as the averages of the corresponding tristimulus values of the spots, thus:

$$X_a = (X_1 + X_2 + X_3 + \dots + X_n)/n \quad (3)$$

and similarly for Y_a and Z_a .

The tristimulus value of the light transmitted by three layers of colorants (say magenta, lemon yellow, and cyan) are integrals like those of Eq. (1), where the spectral radiance distribution $dL/d\lambda$ of the source is supplanted by: $T_m T_{ly} T_c(dL/d\lambda)$, where T_m, T_{ly} , and T_c are the spectral transmittances of the magenta, the lemon yellow, and the cyan layers, respectively.

Theories of Color. It is generally supposed that the incidence of radiant energy on an element of the retina is detected by four photopigments present there; one of these (rhodopsin) serves for twilight vision yielding gray perceptions (see curve 6, Fig. 1); the other three (chemical composition unknown) in combination, serve for daylight vision yielding chromatic perceptions (see curves 1, 5, and 7).

Rhodopsin is found in the retinal rods (duplicity theory). The other three photopigments are found in the retinal cones, either segregated full strength (curves 5 and 7), segregated at low concentration

(curve 2), or integrated (either curve 3 or curve 4). Normal vision is explained by triads of receptors (curves 3, 5, 7, or curves 4, 5, 7) according to the three-components or Young-Helmholtz theory.

The signals from the retinal cones are processed on their way to the visual center of the cortex in the occipital lobe of the brain. It is generally supposed that the red, green, and blue signals leaving the retinal cones are converted by this processing into light-dark, yellow-blue, and red-green signals at the occipital lobe. These processes correlate with the following functions of the tristimulus values: *Y* (light-dark), *Y-Z* (yellow-blue), and *X-Y* (red-green). The explanation of normal color vision in terms of these processes is the opponent-colors, or Hering, theory of vision.

Explanations of color vision in terms of a photopigment stage, a receptor stage, and an optic-nerve or cortical stage, are called stage theories of vision (Müller, Adams).

DEANE B. JUDD

References

Evans, Ralph M., "An Introduction to Color," New York, John Wiley & Sons, 1948.
 LeGrand, Y., "Light, Colour and Vision," New York, John Wiley & Sons, 1957.
 Judd, Deane B., and Wyszecki, Günter, "Color in Business, Science, and Industry," New York, John Wiley & Sons, 1963.
 Wright, W. D., "The Measurement of Colour," London, Hilger & Watts, 1963.

Cross-references: LIGHT, VISION AND THE EYE.

COLOR CENTERS

The term "color center" is broadly used to describe those microscopic defects that produce a change in the optical transparency of materials. The prototypes of these defects are found in the crystals of alkali halides. These crystals consist of positive alkali ions, e.g., K^+ , and negative halide ions, e.g., Cl^- , which are arranged alternately in

a three-dimensional array. If a negative halide ion is removed from its lattice site and put at some other position in the crystal, such as on the surface or into the space between other normal ions, thereby forming an "interstitial ion," a negative ion vacancy results. An electron can be trapped by the positive electric field that arises from the positive ions that surround the negative ion vacancy, thereby forming what may be crudely considered as the analog of the hydrogen atom with the positive field arising from the surrounding alkali ions replacing the positive field of the nucleus. This crystal defect—an electron trapped at a negative ion vacancy—is called an "F-center." The F-center may be in its lowest energy state—the ground state—or it may be excited to higher states by the absorption of a photon of a characteristic frequency. The electron interacts strongly with the positive ions which neighbor the vacancy. Since the position of these positive ions determines the strength of this interaction, any motion of these ions alters the characteristic frequency which is necessary to excite the F-center. There is, therefore, a "band" of frequencies that can be absorbed by a crystal containing F-centers. As the temperature of the lattice is decreased, the motion of the atoms decreases—thereby decreasing the width of the band of frequencies which can be absorbed. Cooling below the temperature of liquid nitrogen ($-196^\circ C$) results in little further change in the width of the band since the vibration of the atoms in the lattice is almost entirely a result of zero-point vibrations below this temperature. As can be seen in Table I, the wavelength (or frequency) of the maximum of the absorption band is characteristic of F-centers in a particular alkali halide crystal.

Following the absorption of a photon, the center is in an excited state. It can return to its ground state by luminescing (i.e., emitting a photon), by ionization of the electron thereby freeing it from the defect, or by a non-radiative process in which the excess energy appears as heat. The relative probability of each of these processes depends upon such parameters as the

TABLE I. COLOR CENTERS CHARACTERISTIC IN SOME ALKALI HALIDES

	Wavelength of the Maximum of the Absorption Bands (Angstroms)						
	F	M	R ₁	R ₂	V _k	H	α
<i>KCl</i>							
20 C	5600	8220	6800	7400	—	—	
-196 C	5400	8010	6590	7290	—	—	1770
-269 C	5370	7980	—	—	3650	3350	
<i>LiF</i>							
20 C	2490	4470	3100	3800	3480	—	
<i>KI</i>							
-269°C	6590	10100	8100	9050	4040	—	2380
<i>RbBr</i>							
-196 C	6770	9570	8050	8590	—	—	

temperature, the concentration, and the types of defects. At low temperatures and for low concentrations of defects, the probability that the F-center luminesces approaches unity. As the temperature is raised, the probability for luminescence decreases while the probability for ionization increases. A band of frequencies is emitted when excited F-centers luminesce. The principal frequency of the emitted radiation is always less than the principal frequency of the band of absorbed frequencies. Since the electron distribution in the excited state is different from that in the ground state, the ions in the neighborhood of the F-center shift to a new equilibrium position in the excited state following the absorption of a photon. Energy is therefore given to the lattice with the resulting emitted photon having an energy less than that of the absorbed photon. This is called the Stokes shift of the luminescence.

The ionization of the electron may be detected by observing the current that flows in the presence of an electric field when light is absorbed by the F-center. The disappearance of this photoconductivity at low temperatures is evidence that the first excited state of the center is bound. That is, the electron can be freed from, or ionized out of, the first excited state only if additional energy is provided by the thermal vibrations of the lattice. It is also observed that a strong electric field can provide sufficient energy to "field ionize" the electron out of the first excited state.

The microscopic models of a number of the prominent electron excess centers in the alkali halides are as follows: *F-center*—an electron trapped at a negative ion vacancy. The principal optical absorption band arises from an electronic transition that may be qualitatively classified as $1s \rightarrow 2p$. *M-center*—two F-centers which occupy nearest neighbor positions. The center resembles a hydrogen molecule. The principal optical absorption band corresponds to a ${}^1\Sigma_g \rightarrow {}^1\Sigma_u$ transition. *R-center*—three F-centers that occupy nearest-neighbor positions in a plane. Two distinct absorption bands are associated with this defect. *F'-center*—two electrons trapped at a negative ion vacancy, i.e., an F-center that has trapped an electron.

Alkali halide crystals exhibit strong optical absorptions in the ultraviolet region of the spectrum. The lowest energy of these transitions can be visualized as arising from the transfer of an electron from a negative halide ion to a positive alkali ion thus producing what is called an "exciton." The energy of this transition is perturbed by crystal defects and gives rise to the so-called Greek absorption bands. The α -band arises from an exciton transition in the neighborhood of a negative ion vacancy. The β -band arises from an exciton transition in the neighborhood of an F-center.

In addition to the defects that have trapped an electron, there also exist defects that have a deficiency of electrons—the so-called hole traps. The *V_K-center* consists of a hole that is shared by two negative halide ions. No vacancies are involved in this center. It may be pictured as a Cl_2^-

molecule that occupies two normal halide ion lattice sites in the crystal. The *H-center* consists of a hole that is shared by four negative halide ions that occupy three normal halide ion lattice sites in the crystal. Although this center is also molecular in nature, it may be conveniently pictured as an interstitial halide atom.

Impurities, such as Ti^{2+} , Pb^{2+} , Ag^+ and Cu^+ , introduce one or more absorption bands into the alkali halides when they are substituted for the alkali ion. The nature of these absorption bands depends upon the specific impurity. Impurities can also modify the properties of some of the color centers described above. If a lithium ion replaces a potassium ion that is the nearest neighbor of an F-center in KCl, it is found that two distinct absorption bands appear instead of the single absorption band that arises from the F-center in the pure crystal. This defect—an F-center with a foreign alkali ion impurity as a nearest neighbor—is called the *F_A-center*.

Color centers may be readily generated by the following three techniques:

(1) Irradiation with x-rays, fast electrons, neutrons, or high-energy protons generates color centers which have trapped electrons in equal numbers with those that have trapped holes. The temperature at which the irradiation is made determines the defects that are produced. At room temperature, F-, M- and R-centers are formed. At -196°C , F-, F', V_K - and α -centers are formed. At the temperature of liquid helium (-269°C), F-, H-, V_K - and α -centers are formed.

(2) Heating of a crystal to several hundred degrees centigrade in an atmosphere of the alkali metal produces crystals with a stoichiometric excess of alkali metal. If the crystal is rapidly cooled from this temperature, the F-center is the prominent defect. Absorption of light by the F-centers generates M- and R-centers if the optical irradiation is made at room temperature and F'-centers if the crystal is at about -100°C .

(3) Passage of a dc electrical current through samples held at several hundred degrees centigrade generates color centers. F-centers are generated at the cathode and move into the crystal under the action of the applied electric field.

Many of the properties of the host crystal are strongly affected by the presence of color centers. The density decreases when F-centers are introduced, the thermal conductivity at low temperatures is reduced, the crystal exhibits paramagnetic properties, and an enhanced electrical conduction is observed when the crystal is irradiated with light. Irradiation with x-rays enhances the hardness of the crystal.

Although the microscopic models of the color centers and their interactions with the host lattice and with other defects are best understood in the alkali halides, a considerable amount of information is available on similar defects in other materials. In many cases, the defects have been introduced by irradiation. Studies have been particularly intensive on the alkaline earth halides and oxides, crystalline quartz, fused silica, aluminum oxide, and diamond. Recent studies of

radiation-induced defects in semiconductors, particularly germanium and silicon, have revealed defects that may be properly classified as color centers.

W. DALE COMPTON

References

Books

- Schulman, J. H., and Compton, W. D., "Color Centers in Solids," London, Pergamon Press, 1962.
Mott, N. F., and Gurney, R. W., "Electronic Processes in Ionic Crystals," Oxford, Clarendon Press, 1940.
Przibram, K., "Irradiation Colours and Luminescence," London, Pergamon Press, 1956.
"Radiation Effects in Inorganic Solids," *Discussions Faraday Soc.*, **31** (1961).

Review Articles

- Seitz, F., *Rev. Mod. Phys.*, **18**, 384 (1946); **26**, 7 (1954).
Compton, W. D., and Rabin, H., *Solid State Phys.*, **16**, 121 (1964).
Pick, H., *Nuovo Cimento*, **7** (2), 498 (1958).

Cross-references: EXCITONS; LUMINESCENCE; PHOTOCONDUCTIVITY; RADIATION, IONIZING, BASIC INTERACTIONS.

COMPRESSIBILITY, GAS

(a) The *compressibility* of a gas is defined as the rate of volume decrease with increasing pressure, per unit volume of the gas. The compressibility depends not only on the state of the gas, but also on the conditions under which the compression is achieved. Thus, if the temperature is kept constant during compression, the compressibility so defined is called the isothermal compressibility β_T :

$$\beta_T = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T = \frac{1}{\rho} \left(\frac{\partial \rho}{\partial P} \right)_T \quad (1)$$

If the compression is carried out reversibly without heat exchange with the surroundings, the *adiabatic compressibility* at constant entropy, β_S , is obtained:

$$\beta_S = \frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_S = \frac{1}{\rho} \left(\frac{\partial \rho}{\partial P} \right)_S \quad (2)$$

Here P is the pressure, V the volume, ρ the density, T the temperature, and S the entropy.

In adiabatic compression, the temperature rises, thus the pressure increases more sharply than in isothermal compression. Therefore β_S is always smaller than β_T .

The *compressibility factor* of a gas is the ratio PV/RT (cf. GAS LAWS). This name is not well chosen since the value of the compressibility factor by itself does not indicate the compressibility of the gas.

(b) *Experimental values* for the compressibility of gases can be obtained in several ways, most of which are indirect.

Since the compressibility is proportional to the pressure derivative of the volume, any experiment that establishes the P - V - T relation of a gas with sufficient accuracy also yields data for the isothermal compressibility. For obtaining the adiabatic compressibility from the P - V - T relation some additional information is necessary [see section (c)], for instance SPECIFIC HEAT data in the perfect gas state of the substance considered. For recent reviews of experimental methods for determining P - V - T relations of gases, see references 1, 2a and 2b. A more direct way of determining the adiabatic compressibility is by measuring the speed of sound v , the two quantities being related by

$$v^2 = \frac{1}{\rho \beta_S} \quad (3)$$

This relation is valid only when the compressions and expansions of the sound wave are truly reversible and adiabatic. This is the case if the frequency is fairly low and the amplitude small. Experimental techniques for determining the speed of sound are discussed in the literature.^{2a,3}

(c) The experimental behavior of the compressibility as a function of pressure and temperature is as follows: Dilute gases obey the laws of Boyle and Gay-Lussac, $PV = RT$, to a good approximation. Then it is readily shown that the following relations hold for the compressibility:

$$\begin{aligned} \beta_T &= 1/P - V/RT \\ \beta_S &= 1/\gamma P - V/\gamma RT \end{aligned} \quad (4)$$

where $\gamma = c_p/c_v$, the ratio of the specific heats at constant volume and at constant pressure, respectively, and R is the gas constant.

Compressed gases show large deviations from the behavior predicted by Eq. (4). This is demonstrated in Fig. 1, where the isothermal compressibility of argon, divided by the corresponding value for a perfect gas at the same density, is pictured as a function of density for various temperatures. It is seen, first of all, that at all temperatures the compressibility at high densities falls to a small fraction of the value for a perfect gas, and secondly, that super-critical isotherms show a maximum in the ratio $\beta/\beta_{\text{perfect}}$ as a function of density, which maximum is the more pronounced the closer the critical temperature. It occurs roughly at the critical density ρ_c . Since at the critical point $(\partial P/\partial V)_T$ equals zero, the isothermal compressibility becomes infinite at this point. The adiabatic compressibility, however, remains finite. Qualitatively, all gases show the same behavior as pictured for argon in Fig. 1.

(d) The *molecular theory* can explain the general features of the compressibility in its temperature and density dependence. The pressure of the gas is caused by the impact of the molecules on the wall. If the volume is decreased at constant temperature, the average molecular speed and force of impact remain constant, but the number of collisions per unit area increases and thus the

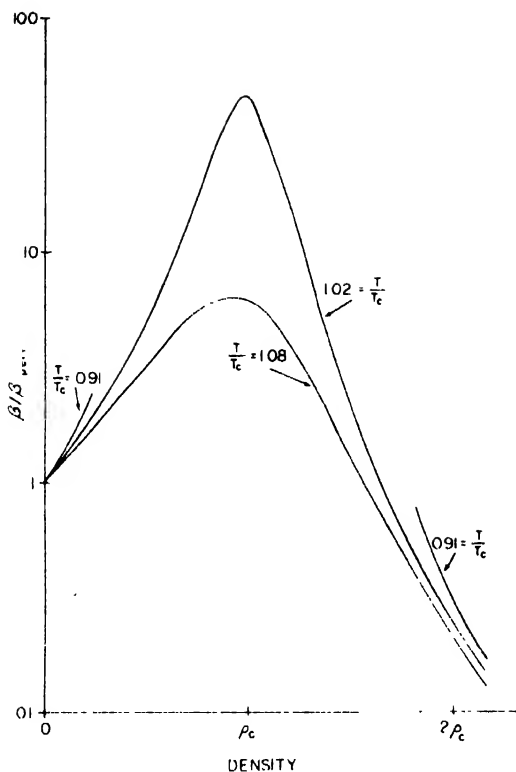


FIG. 1. The ratio $\beta/\beta_{\text{perf.}}$ of the isothermal compressibility of argon to that of a perfect gas at the same density, as a function of the density, at 0.91, 1.02 and 1.08 times the critical temperature. The critical density is indicated by ρ_c .

pressure rises. If the gas is compressed adiabatically, the heat of compression cannot flow off, thus the average molecular speed and force of impact increase as well, giving rise to an extra increase of pressure. Therefore $\beta_s > \beta_T$. The actual magnitude of the temperature rise depends on the internal state of the molecules: the more internal degrees of freedom available, the more energy can be taken up inside the molecule and the smaller the temperature rise on adiabatic compression. Thus for gases consisting of molecules with many internal degrees of freedom, adiabatic and isothermal compressibilities differ but little.

If the gas is assumed to consist of molecules of negligible size and without interaction, then the gas can be shown to follow the laws of Boyle and Gay-Lussac, therefore its isothermal and adiabatic compressibilities must be given by Eq. (4). For a perfect gas, the percentage pressure rise is proportional to the percentage volume decrease if the change is small; thus the compressibility is inversely proportional to the pressure.

To explain the very different behavior of real gases, the model must be modified. Suppose the molecular volume is small but not negligible. In

states of high compression, where the total molecular volume becomes of the order of the volume available to the gas, the free space available to the molecules is only a fraction of what it would be in a perfect gas, and thus the real gas is much harder to compress than the perfect gas. This explains the low compressibility of dense gases and liquids (Fig. 1).

Furthermore, one assumes that molecules, on approaching each other, experience a mutual attraction before they collide; this mutual attraction makes it easier to compress a real gas than a perfect gas. This explains the initial rise of the compressibility of a real gas over that of a perfect gas (Fig. 1) at temperatures not too far above the critical.

When compressed at subcritical temperatures, the gas condenses; that is, macroscopic clusters or droplets are formed under the influence of the attractive forces. During condensation the pressure remains constant while the volume decreases, giving rise to an infinite compressibility in the two-phase region. At the critical point the system is on the verge of condensation and the compressibility is also infinite.

(e) *Theoretical predictions* for the isothermal compressibility can obviously be obtained from any theory of the equation of state (cf. GAS LAWS). If, in addition, data for the specific heat are supplied, the adiabatic compressibility can be derived in the same way. Thus the compressibility can be derived from the virial expansion of the equation of state which expresses the ratio PV/RT in a power series in the density, the coefficients being related to the interactions of groups of two, three, etc., particles.^{4,5}

In the dense system the convergence of the virial expansion is doubtful. In any case the higher coefficients are hard to calculate; here approximate theories have been developed, of which the cell model⁵ is an example.

Many semiempirical equations of state with varying degree of theoretical foundations are in use. The van der Waals equation, a two-parameter equation which gives a qualitatively correct picture of the P - V - T relations of a gas and of the gas-liquid transition, is an example. For a survey of the most useful semiempirical equations of state, see reference 5.

Modern developments are centered around the calculation of the radial distribution function $g(r)$, which is the ratio of the density of molecules at a distance r from a given molecule, to the average density in the gas. The compressibility can be expressed straightforwardly in terms of $g(r)$ as follows:⁵

$$KT\beta_T = 1/\rho + \int_0^\infty [g(r) - 1]4\pi r^2 dr \quad (5)$$

Approximate evaluations of the radial distribution function in dense systems are being obtained as solutions to integral equations derived from first principles under well-defined approximations. For a survey of the approaches used, see reference 6.

J. M. H. LEVELT SENGERS

References

1. Rowlinson, J. S., "The Properties of Real Gases," in Flügge, S., Ed., "Handbuch der Physik," Vol. 12, p. 1, Berlin, Springer-Verlag, 1958.
2. van Itterbeek, A., Ed., "Physics of High Pressure and the Condensed Phase," Amsterdam, North Holland Publishing Company, in press.
 - a. Levelt Sengers, J. M. H., "The Experimental Determination of the Equation of State of Gases and Liquids at Low Temperatures."
 - b. Verbeke, O., and van Itterbeek, A., "Density Measurements on Liquefied Gases."
 - c. van Dael, W., and van Itterbeek, A., "Sound Velocity Measurements in Fluids and Gases under High Pressure."
3. Barone, A., "Generation, Detection and Measurement of Ultrasound," in Flügge, S., Ed., "Handbuch der Physik," Vol. 11, 2, p. 74, Berlin, Springer Verlag, 1961.
4. Mayer, J. E., and Mayer, M. G., "Statistical Mechanics," Ch. 13, New York, John Wiley & Sons, 1940.
5. Hirschfelder, J. O., Curtiss, C. F., and Bird, R. B., "Molecular Theory of Gases and Liquids," New York, John Wiley & Sons, 1954.
6. Rushbrooke, G. S., "Distribution Function Theories of Fluids," Oxford, Pergamon Press, in press.

Cross-references: GAS LAWS, KINETIC THEORY.

COMPTON EFFECT

Introduction. The Compton effect refers to the collision of a photon and a free electron in which the electron recoils and a photon of longer wavelength is emitted as indicated in Fig. 1. It is one of the most important processes by which x-rays and γ -rays interact with matter and is also one which is accurately calculable theoretically.

A discussion of the effect is found in most textbooks on atomic physics. Particularly complete presentations have been made by Evans.^{1,2}

History. Barkla and others (1908) made many observations on the scattering of x-rays by different materials. The diffuse scattering was interpreted qualitatively by J. J. Thomson in terms of the interaction of electromagnetic waves with electrons which he had shown to be a constituent of all atoms. As more experiments were carried out with light elements, it was established by J. A. Gray (1920) that the diffusely scattered x-rays were less penetrating. This implied that the scattered radiation had a longer wavelength than the incident radiation. This could not be reconciled with Thomson's theory which represented x-rays as continuous electromagnetic waves with wavelengths unchanged by scattering.

The effect which now bears his name was established quantitatively by Arthur Holly Compton (1923) when he published careful spectroscopic measurements of x-rays scattered at various angles by light elements. He found that x-rays scattered at larger angles had systematically larger wavelengths. In searching for an explanation of the data, he discovered that the observations were accounted for by considering the

scattering as a collision between a single photon and a single electron in which energy and momentum are conserved.

The important place which the effect occupies in the development of physics lies in his interpretation of the effect in terms of the newly emerging quantum theory. The essential duality of waves and particles was demonstrated in an especially clear way, since the collision conserved energy and momentum while both the incident and scattered x-rays revealed wave-like properties by their scattering from a crystal. In recognition for this contribution, Compton was awarded the Nobel Prize in 1927.

A complete theory for the effect was worked out in 1928 by Klein and Nishina using Dirac's relativistic theory of the electron. The calculation was one of the brilliant successes of the Dirac theory. It represents quantitatively, within the experimental uncertainties, all phenomena associated with the scattering of photons by electrons for energies up to several billion electron volts. Because of the confidence with which photon interaction with electrons can be interpreted, the Compton effect has been important in the analysis of the energy and the polarization of gamma rays from many sources.

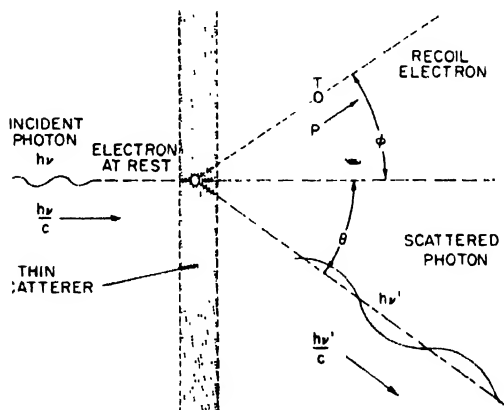


FIG. 1. Diagram showing the initial and final energies and momenta for Compton scattering.

Kinematics. The relations between the energies and directions of the incident and scattered photons and the recoil electron are determined by the conservation of energy and of the components of momentum parallel and at right angles to the incident beam. In the usual case, where the electron is initially at rest and the energy and momentum of the incident photon are $h\nu$ and $(h\nu/c)$, the equations are:

$$h\nu = h\nu' + T \quad (1)$$

$$\frac{h\nu}{c} = \frac{h\nu'}{c} \cos \theta + p \cos \phi \quad (2)$$

$$0 = \frac{h\nu'}{c} \sin \theta - p \sin \phi \quad (3)$$

where c is the velocity of light, h is Planck's constant, and the angles are those indicated in Fig. 1. The relativistic relation between the kinetic energy T of the recoiling electron and its momentum p is

$$pc = \sqrt{T(T + 2mc^2)} \quad (4)$$

where m is the mass of the electron. These equations can be combined to obtain relations which are useful in the interpretation of data. The Compton shift is

$$\lambda' - \lambda = \frac{c}{\nu'} - \frac{c}{\nu} = \frac{h}{mc} (1 - \cos \theta) \quad (5)$$

This relation was first found experimentally by Compton, who noted that the shift in wavelength ($\lambda' - \lambda$) depended on the angle, but not on the wavelength, of the incident photon. The quantity (h/mc), which is the shift at 90° , is called the Compton wavelength of the electron and is one of the useful constants (2.4262×10^{-10} cm).

$$h\nu' = \frac{mc^2}{1 - \cos \theta + \frac{mc^2}{h\nu}} \quad (6)$$

In this form, the energy of the scattered photon is seen to vary from that of the incident photon at 0° to less than $(mc^2/2)$ at 180° . At high energies the angle θ for which $h\nu'$ is $(h\nu/2)$ is approximately $2(mc^2/h\nu)$ radians.

The kinetic energy of the recoiling electron is

$$T = \frac{h\nu(1 - \cos \theta)}{(1 - \cos \theta) + \frac{mc^2}{h\nu}} \quad (7)$$

The relation between the scattering angles of the electron and photon is

$$\cot \phi = \left(1 + \frac{h\nu}{mc^2}\right) \left(\frac{1 - \cos \theta}{\sin \theta}\right) \quad (8)$$

Graphs of these kinematic relations and of the scattering cross section are given by Evans⁽²⁾ and by Nelms⁽³⁾.

Scattering of Unpolarized Radiation. The differential cross section for the scattering of unpolarized radiation at an angle θ is given by the Klein and Nishina equation.

$$\frac{d\sigma}{d\Omega} = \frac{r_0^2}{2} \left(\frac{\nu'}{\nu}\right)^2 \left(\frac{\nu}{\nu'} + \frac{\nu'}{\nu} - \sin^2 \theta\right) \quad (9)$$

where r_0 is the electron radius $= e^2/mc^2 = 2.8177 \times 10^{-13}$ cm, and ν' is obtained from Eq. (6). The cross section is shown as a function of θ for several energies in Fig. 2. The classical Thomson cross section $r_0^2 (1 + \cos^2 \theta)/2$ can be seen to hold for low energies where $\nu' \approx \nu$.

The total cross section obtained by integrating this cross section over angle is important in the attenuation of well-defined beams in passing through a material. The relative importance of

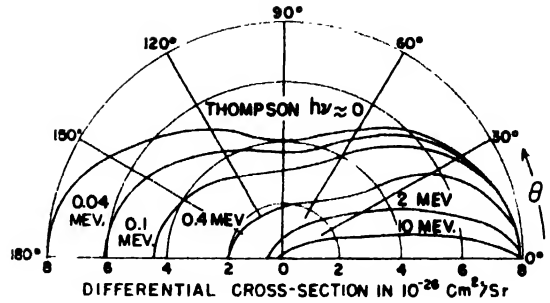


FIG. 2. Differential cross section for photons scattered at angles, θ , for a number of incident energies.

Compton scattering as compared to the photoelectric effect and pair production is illustrated for aluminum in Fig. 3 where the attenuation coefficient α is shown as a function of energy. The fraction of the photons surviving without an interaction upon passing through x g/cm² of aluminum is $e^{-\alpha x}$. The Compton effect is the major one between 0.5 and 2 MeV. Extensive tables and graphs for other elements are available.^{2,1}

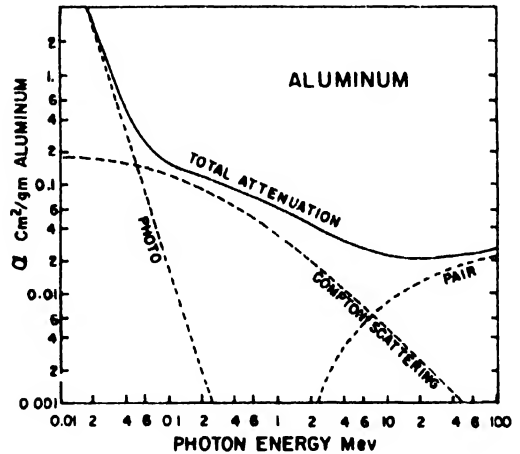


FIG. 3. The attenuation coefficients, α , for the absorption of photons in aluminum as a function of energy. The broken lines represent the separate contributions of the photoelectric effect, the Compton effect, and pair production to the absorption.

In detectors whose response is proportional to the energy deposited by the recoil electrons, the distribution of electron energies associated with a photon of known energy is of interest. The distribution is given by the relation

$$\frac{d\sigma}{dT} = \frac{\pi r_0^2 mc^2}{(h\nu)^2} \left\{ 2 + \left(\frac{T}{h\nu - T} \right)^2 \left[\frac{(mc^2)^2}{(h\nu)^2} + \frac{h\nu}{T} - \frac{2mc^2(h\nu - T)}{h\nu T} \right] \right\}$$

where T varies from 0 to $T_{\max} = 2(h\nu)^2/(2h\nu + mc^2)$. A number of these distributions are shown in Fig. 4.

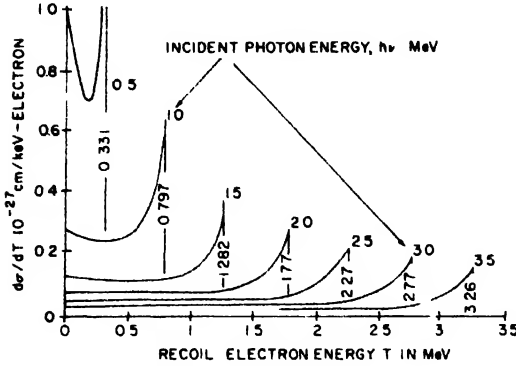


FIG. 4. The energy distribution of the Compton recoil electrons for several values of the incident photon energy $h\nu$. Based on figure in "Compton Effect" by R. D. Evans in "Handbuch der Physik," Vol. XXXIV, pp. 234-298, 1958, J. Fluge, Ed., by permission of Springer-Verlag, publishers.²

Scattering of Plane Polarized Radiation. The differential cross section for the scattering of plane polarized radiation by unoriented electrons was also derived by Klein and Nishina. It represents the probability that a photon, passing through a target containing one electron per square centimeter, will be scattered at an angle θ into a solid angle $d\Omega$ in a plane making an angle η with respect to the plane containing the electric vector of the incident wave.

$$\frac{d\sigma}{d\Omega} = \frac{r_0^2}{2} \left(\frac{\nu'}{\nu} \right)^2 \left(\frac{\nu}{\nu'} + \frac{\nu'}{\nu} - 2 \sin^2 \theta \cos^2 \eta \right) \quad (9)$$

The cross section has its maximum value for $\eta = 90^\circ$, indicating that the photon and electron tend to be scattered at right angles to the electric vector of the incident radiation.

This dependence is the basis of several instruments for determining the polarization of photons. For example, it was used by Wu and Shaknov⁴ to establish the crossed polarization of the two photons emitted upon the annihilation of a positron electron pair; by Metzger and Deutsch⁵ to measure the polarization of nuclear gamma rays; and by Motz⁶ to study the polarization of bremsstrahlung.

Scattering of Circularly Polarized Radiation. The scattering of circularly polarized photons by electrons with spins aligned in the direction of the incident photon is represented by

$$\frac{d\sigma}{d\Omega} = r_0^2 \left(\frac{\nu'}{\nu} \right)^2 \left[\left(\frac{\nu}{\nu'} + \frac{\nu'}{\nu} - \sin^2 \theta \right) \pm \left(\frac{\nu}{\nu'} - \frac{\nu'}{\nu} \right) \cos \theta \right] \quad (10)$$

The first term is the usual Klein-Nishina formula for unpolarized radiation. The $+$ sign for the additional term applies to right circularly polarized photons. The ratio of the second term to the first is a measure of the sensitivity of the scattering as a detector of circularly polarized radiation and is shown in Fig. 5.

In practice, the only source of polarized electrons has been magnetized iron where 2 of the 26 electron spins can be reversed upon changing its magnetization. Although the change in the absorption or scattering is usually only a few per cent, this is often sufficient to get accurate and reliable measurements of circular polarization.

Cross sections for some practical arrangements and discussions of earlier work are presented by Tolhoek⁸. Applications to the determination of the helicities of photons, electrons, and neutrinos in confirming the two-component theory of the neutrino are reviewed in considerable detail by L. Grodzins⁹.

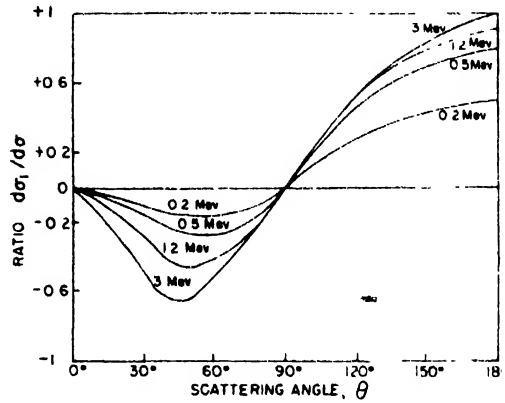


FIG. 5. The ratio of the partial cross section dependent on the spin orientation of the electrons to the average cross section as represented by the second and first terms of Eq. (10). Based on figure in "Compton Effect" by R. D. Evans in *Handbuch der Physik*, Vol. XXXIV, pp. 234-298, 1958, J. Fluge, Ed., by permission of Springer-Verlag, publishers.²

Proton and Deuteron Compton Effect. Particle-like scattering of high-energy photons by protons and deuterons has been observed and has been referred to as the proton and deuteron Compton effect. The kinematic equations are identical to those for electrons except that the mass is that of the proton or deuteron.

Although the cross sections are smaller than that for electrons, by the square of the ratio of the masses, the scattering is easily distinguished by the characteristically higher energy of the radiation at large angles. At energies above the pion threshold, the cross section is dominated by pion nucleon resonances. The experimental cross sections for the scattering by protons, as presented by Steining, Loh, and Deutsch,¹⁰ are shown in Fig. 6. Some experimental results and

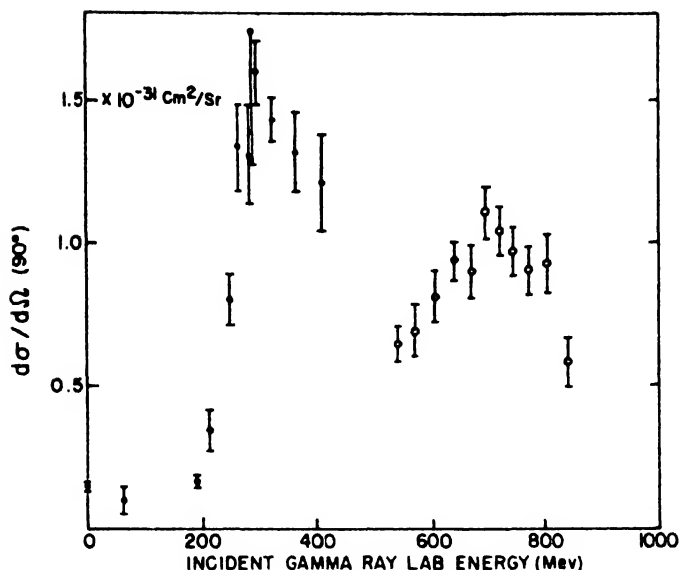


FIG. 6. The differential cross section for the scattering of high-energy photons by protons at 90° in the center of mass system. Based on figure in Steining, Loh and Deutsch, *Phys. Rev. Letters*, **10**, 536 (1963).¹⁰

calculations on the coherent scattering from deuterium are described by Jones, Gerber, Hanson, and Wattenberg¹¹.

Cross-references: COLLISIONS OF PARTICLES, CONSERVATION LAWS AND SYMMETRY, QUANTUM THEORY.

A. O. HANSON

COMPUTERS

References

1. Evans, R. D., "The Atomic Nucleus," Chapter 23 New York, McGraw-Hill Book Co., 1955.
2. Evans, R. D., "Compton Effect," In Flugge, S. Ed., "The Encyclopedia of Physics" Vol. 34, pp 234-298, Berlin, Springer-Verlag, 1958.
3. Nelms, A. T., "Graphs of the Compton Energy-Angle Relationship and the Klein-Nishina Formula from 10 keV to 500 MeV," *Natl. Bur. Std. Circ.*, **542** (1953).
4. White, G. R., "X-Ray Attenuation Coefficients from 10 keV to 100 MeV," *Natl. Bur. Std. Rept.*, **1003** (1952).
5. Wu, C. S., and Shaknov, I., "Angular Correlation of Scattered Annihilation Radiation," *Phys. Rev.*, **77**, 136 (1950).
6. Metzger, F., and Deutsch, M., *Phys. Rev.*, **78**, 551 (1950).
7. Motz, J. W., "Bremsstrahlung Polarization Measurements for 1 MeV Electrons," *Phys. Rev.*, **104**, 557 (1956).
8. Tolhoek, H. A., "Electron Polarization, Theory and Experiment," *Rev. Mod. Phys.*, **28**, 277 (1956).
9. Grodzins, L., "Measurement of Helicity," in Frisch, O. R., Ed., "Progress in Nuclear Physics," New York, Pergamon Press, 1959.
10. Steining, R. F., Loh, E., and Deutsch, M., "The Elastic Scattering of Gamma Rays by Protons," *Phys. Rev. Letters*, **10**, 536 (1963).
11. Jones, R. S., Gerber, H. J., Hanson, A. O., and Wattenberg, A., "Deuteron Compton Effect," *Phys. Rev.*, **128**, 1357 (1962).

Computers may be mechanical, electromechanical, or electronic devices. They may operate in the analog or the digital mode, or a combination of the two. Whatever their form may be, their basic function is the automatic performance of a selected set of mathematical or logical operations.

Analog Computers. The analog computer (sometimes called the continuous computer), is often identifiable as a physical model of the problem to be solved. Problem data are represented by physical quantities like lengths, voltages, fluid densities, etc., and solutions are obtained through measurement of output quantities. A simple illustration is the slide rule which adds lengths representing the logarithms of the factors to give as the measure of a product a length representing its logarithm. Other illustrations are artificial transmission lines used to predict the performance of proposed electric power transmission lines, and electrical linear equation solvers. The former is a reduced scale model of the proposed power line; the latter is a model of the system of equations which consists of coupled circuits to represent the equations, electrical components to introduce the coefficients, and electrical meters to indicate the solution.

An analog computer may be a large device containing components which represent elementary mathematical functions and components which perform integration or differentiation. Devices of this type, called differential analyzers, have been used extensively for the solution of differential

equations. The first differential analyzer was designed by Vannevar Bush about 1930 and consisted of mechanical integrators of the wheel-and-disk type, a frame containing removable shafts, changeable gears and couplings, input handwheels and recording pens for producing graphs of problem solutions. Later Bush, Caldwell, and others developed more automatic differential analyzers in which the coupling between integrators was controlled by switching techniques enabling users to program on punched paper tape the interconnections of components to set the device up for problems. Still later the mechanical integrators gave way to electric integrators, and compact differential analyzers were developed consisting entirely of electrical circuitry and meters, with the exception of the input units.

Sophisticated computers have been built whose basic operation is straightforward counting. One such computer, the Maddida, appeared in 1951. It performed mathematical operations by reducing them to counts. It was characterized as a digital differential analyzer and was applied to the solution of differential equations. Its precision was higher than that within the realm of capability of analog computers.

Digital Computers. The digital computer, as its name implies, is an extension of finger counting. It is as old as the abacus, which has been in use for at least twenty-five centuries. And forerunners of the modern digital desk calculating machine were presented to the world by Blaise Pascal in France in 1642 and Gottfried Wilhelm Leibnitz of Germany in 1671. However, current emphasis on computer development and use is centered about a recently developed device, the automatically sequenced electronic digital computer.

In 1786, J. H. Müller, an engineer, conceived the idea of an automatic digital computer, but he was discouraged by the technical difficulties barring its development. About twenty-five years later, Charles Babbage, a young English mathematician, made suggestions for an automatic digital "difference engine" designed to build up mathematical functions automatically by the use of differences and to print the answers without human intervention, except to start the computation. Before completing this device, Babbage turned to a more general concept, called the "analytical engine," which was logically similar to the modern automatic digital computer. The analytical engine was not completed, although Babbage worked on it almost half a century until his death. The design requirements were beyond the purely mechanical techniques of his day.

A limited realization of Babbage's dream of a universal computer took place at the turn of the century, with the invention of punched-card methods in 1890 by Herman Hollerith. Electromagnetic techniques had been perfected which made possible the realization of Hollerith's ideas and led to the development of commercial punched-card accounting equipment of wide-

spread application. In this computing equipment data are represented by holes in punched cards and the sequence of operations is determined by the configuration of patch-cords in electrical plug boards.

Work was begun independently by both Howard Aiken and George Stibitz on automatic relay computers in about 1937. The ensuing computers embodied the dream of Babbage that was beyond the technical facilities available in his day—controls that provided for the automatic sequencing by punched cards or paper tape of all the operations required for the solution of a problem, including the acquisition of data, the calculations, and the recording of results. Some relay computers even had the ability to discontinue a problem and begin working on another, according to criteria introduced into the machine with the set of instructions, or program, inserted initially into the device.

The development of pulsed electric circuitry for RADARS made possible the utilization of the fast response of electronic tubes in computation. The first electronic digital computer, the Eniac (Electronic Numerical Integrator and Calculator) was developed under the leadership of John Mauchly and J. P. Eckert. Even though radar technology was then available, the use of electron tubes as relays was novel, and the reliability requirements on the electronic counting and switching circuits in the Eniac exceeded those of any prior application.

While the Eniac was an extremely fast computer, operating at the speed of electronic circuitry, it was sequenced through the setting and placement of hundreds of switches and electrical jumpers. Problem set-up time, even for calculations of only moderate complexity, could occupy days or even weeks, during which time the machine was not available for other use. The designers, not satisfied with this feature, were planning a more automatically sequenced electronic digital computer, called Edvac (Electronic Discrete Variable Automatic Computer), even before Eniac's completion. Edvac's design was the forerunner of the modern stored program, automatically sequenced electronic digital computer.

Functionally described, the present electronic digital computer comprises four components: an arithmetic or computing unit, a high-speed internal storage unit, input-output devices (which may be a single dual-purpose unit), and a control unit.

Data and instructions are indistinguishable within the electronic digital computer; both are represented for example by binary patterns of vacuum tube, transistor, or diode states in the electronic circuitry, and by such patterns of magnetized spots on small toroidal cores or thin magnetizable film in the high-speed storage. In the preparation of instruction sequences, called programming, the user must arrange to have the control unit refer to the correct locations in the high-speed storage for instructions. In addition, computers often contain auxiliary storage in the

form of magnetic tape or rotating magnetic drum.

For effective exploitation of its high operating speeds, the electronic digital computer contains built-in branching operations, which enable the control to select alternative paths according to the results of prior calculations. In addition to this flexibility, since instructions and data are indistinguishable in form, the arithmetic unit can modify the stored instructions by performing calculations on them. By employing these features, having received initially a complete set of instructions, electronic digital computers can solve without human intervention problems of tremendous complexity and magnitude. This characteristic has caused the devices to be called electronic brains and, more conservatively, stored-program computers.

The electronic digital computer is normally a high-precision device, capable of handling numbers of twelve-decimal digits or more in coded binary form or straight binary numbers of 35 bits or more. Theoretically the computer can provide whatever precision is required by distributing numbers over more than single registers, or single storage locations, and the use of special, multi-precision programs. In actuality multi-precision techniques slow down the computers to such an extent that rarely does one go beyond double-precision operation.

Numbers may be represented within the computer in logarithmic form, a mantissa together with the exponent of the power of the number base giving the appropriate scale factor. This representation simplifies programming by eliminating the need for insertion of scaling factors in programs to keep data from spilling off the computer registers. The logarithmic representation is called floating point representation. Floating point operation may be programmed on computers which do not have it as a built-in feature, but at significant loss in operating speed.

The set of computer instructions required for the solution of a problem, together with the numerical constants required, is called a "code" or a "routine." Routines may be built up from "subroutines" which are routines for frequently occurring calculations, ranging from the calculation of the trigonometric functions, logarithms, and exponentials to the solution of differential equations.

A commonly used form of computer instruction is the single-address instruction which designates the operation to be performed and the location, or address, of an operand in the internal storage. The second operand may be taken from a selected register, for example an accumulator register in the arithmetic unit. Unless otherwise specified by jump or branching instructions, the single-address computer executes serially the instructions placed in successive storage locations, called addresses.

A section of single address coding is illustrated by the following; where the operations are designated mnemonically, (AC) and (M) designate the contents of the accumulator register and internal storage location M respectively:

Address	Instruction	Operations performed
100	CLA 198	(AC) is replaced by (198), (198) is unaffected.
101	SUB 199	(AC) is replaced by (AC)-(199), which is (198)-(199).
102	TMI 250	If (AC) is negative, control takes next instruction from address 250, otherwise it takes the next instruction from address 103.

Suppose location 199 contains a measure of accuracy reached in a successive approximation process, computed according to the instruction code, and (198) is the relevant prescribed numerical bound, placed in the code by the programmer; and suppose further, that address 250 contains the first instruction of the approximation subroutine. Then this coding arranges for the computer to turn aside from a routine to perform successive approximation unless desired accuracy has been obtained. This illustrates the branching operating, without which automatic computing of only limited scope would be possible.

The main advantage possessed by the electronic digital computer is its high speed of operation. Arithmetic operations and transfer of information between electronic registers or between high-speed storage and the registers takes at most only a few microseconds. For this reason the computer may be designed to use iterative techniques to perform even basic operations, like the extraction of roots. And in successive approximation techniques a formula may be applied literally thousands of times to obtain the required accuracy, at no great cost in time. The development of computational techniques adapted to the characteristics of the high-speed digital computer, including study of the generation and propagation of error, has given rise to a new mathematical field, called numerical analysis.

Programming, the preparation of instruction sequences for electronic computers, has become a profession. The modern computer may have a hundred or more built-in operations, and programming in terms of the basic operations is usually replaced by use of sets of instructions more natural for the human programmer to use, called programming languages. Among the many such programming languages in use, prepared by computer users and manufacturers, three have been proposed as standards: ALGOL and COBOL, for scientific and business applications, respectively, and FORTRAN, devised for general use, but more suitable for engineering and scientific problems than business application. Special programs for conversion of programs written in such languages into sequences of built-in computer operations, called compilers or translators must be provided. Most manufacturers of electronic digital computers supply these programs with their equipment.

E. W. CANNON

References

Stibitz, George R., and Larrivee, Jules A., "Mathematics and Computers," New York, McGraw-Hill Book Co., 1957.

- Richards, R. K., "Arithmetic Operations in Digital Computers, D. Van Nostrand Co., 1955.
- Von Handel, Paul, Ed, "Electronic Computers, Fundamentals, Systems and Applications," Englewood Cliffs, N.J., Prentice-Hall, 1961.
- Alt, Franz L., "Electronic Digital Computers, Their Use in Science and Engineering," New York, Academic Press, 1958.

CONDENSATION

The condensation of a vapor to form a liquid, other amorphous phase or crystal generally occurs by the mechanisms of nucleation and subsequent growth. Nucleation is a thermally activated process which leads to a stable fragment of the condensed phase. In the absence of surfaces of certain condensed phases, reactive foreign molecules or other potent catalysts to the nucleation process, it is usually the slower step and occurs at an appreciable rate only under conditions considerably removed from equilibrium. For the usual case where nucleation catalysts are present, one characterizes the process as heterogeneous nucleation, but if there are no such catalysts at all, one speaks of homogeneous nucleation.

In principle, statistical thermodynamics would appear to offer the most attractive approach to nucleation rate theory. For example, Band¹ and Hill^{1a} have given formal treatments for the equilibrium concentration of clusters of molecules in a vapor. However, the internal partition functions have thus far eluded quantitative evaluation. Further, in the case of metallic systems, one has at present little knowledge of the electronic energy states in clusters containing only a few atoms.

Accordingly, even to the present day, most treatments follow that of Volmer² and co-workers who evaluated the free energy of formation of clusters by ascribing macroscopic thermodynamic properties to them. Thus the free energy of a droplet is described as the sum of a surface term (area times surface tension) and a volume term (volume times the negative bulk free energy change). An attractive feature of this approach is that it permits ready visualization of the origin of the free energy barrier to nucleation in terms of the maximum in the above sum as a function of size. However, a very unattractive aspect is that the calculated size of the critical nucleus, i.e., the cluster size at the top of the free energy barrier, is only about 100 molecules, which leads one to doubt the applicability of macroscopic concepts in the present examples. Nevertheless, following standard methods,³ this spherical-drop model leads directly to a rather simple expression for the rate of homogeneous nucleation of droplets from supersaturated vapor.

The remarkable agreement² of this macroscopic theory with observations of the critical supersaturations for appreciable nucleation rates of various liquids in cloud chambers stood for many years as the basis of our knowledge of nucleation. Then within the past five years, it was

established that most of the earlier data probably represent heterogeneous nucleation on an attenuated and unknown concentration of gaseous ions.⁴ In fact, it now appears that the only critical supersaturation data for homogeneous nucleation of droplets are those of C. T. R. Wilson⁵ and of C. F. Powell⁶ who reported a "fog limit"* for nucleation of water droplets at a supersaturation ratio of about 5.0 (275°K).† Also, further theoretical work⁴ has recently established that the external partition functions had been omitted from the macroscopic theory. When these contributions are included, the theory predicts a critical supersaturation ratio of about 3.0 for water vapor at 275°K, in poor quantitative agreement with Powell's observations.

A considerable amount of work has been done on the heterogeneous nucleation of metal crystals from thermal vapor beams onto substrates.⁴ Most theoretical approaches follow a macroscopic treatment similar to that outlined above, and quantitative agreement with observed critical supersaturations for appreciable nucleation rate is again not good. However, it has been established that in most cases nucleation occurs by the processes of adsorption, surface diffusion, and statistical fluctuation to form the crystalline nuclei. Also, the qualitative relationships between nucleation and epitaxial deposition are thought to have been reasonably well established.

Growth is the process by which the stable nuclei continue to grow and thereby consume the supersaturated vapor. In general, several mechanisms are involved in the growth process, and some of these are thermally activated. However, the free energies of activation are usually low, and hence most growth processes proceed at an appreciable rate even under conditions close to equilibrium where the gross evaporation flux almost equals the gross condensation flux. The first step in growth from the vapor is thought to be adsorption of the impinging molecule. The overwhelming bulk of both experimental and theoretical work⁴ indicates that the impinging atoms or molecules are in most cases thermally accommodated and adsorbed at the surface before being either reevaporated or integrated into the liquid or crystalline structure. In the case of liquids with asymmetrical molecules, Eyring¹ and co-workers have shown that there is an entropy contribution to the free energy of activation which tends to reduce the adsorption efficiency** and hence the condensation coefficient‡ to somewhat below unity.

In the case of liquids, it is thought that the molecular mobility is sufficiently high that the adsorbed molecules are taken almost immediately

* Denoting production of many droplets, i.e., of the order of 10^7 cm^{-3} .

† Meaning the ratio of actual to equilibrium partial pressure of vapor.

** Ratio of molecules which become adsorbed to those which impinge.

‡ Ratio of gross condensation to impinging fluxes.

into the liquid structure. However, the situation is quite different for crystals, whose surfaces are still most conveniently visualized in terms of the original "atomic building block" model of Kossel⁷ and Stranski.⁸ Thus, in the case of certain surfaces of high index, the kinks in the steps of the atomically rough surface provide ready sinks for adsorbed molecules. In fact, experiment shows that such planes grow so rapidly that they quickly eliminate themselves from the crystal growth form, leaving the smoother surfaces of low index. These closely packed planes contain no steps and kinks to serve as sinks for the admolecules diffusing on the surface. Accordingly for a perfect crystal, it is thought that growth can proceed only by nucleation of new monomolecular layers, whose edges provide the sinks, and their lateral propagation. A typical supersaturation ratio for appreciable growth by this mechanism is of the order of 1.5 for molecular substances, and there is a large amount of theoretical and experimental evidence⁹ for the general occurrence of this type of growth from the vapor at high supersaturations.

The fact that real crystals do indeed grow at much lower supersaturation ratios, of the order 1.01, continued to present a theoretical problem for many years. Then in 1949 Burton, Cabrera and Frank¹⁰ showed that certain emergent dislocations of the screw orientation* must provide a source of monomolecular steps for growth at low supersaturations. Further, they demonstrated that the resultant growth form on the close-packed surface, the growth spiral, cannot exterminate itself as do other types of steps or ledges. In the usual case, crystal growth by this mechanism is thought to be controlled by surface diffusion of the admolecules. Experimental verification of these predictions is now voluminous.¹¹

In the interests of brevity, the complex and interesting effects relating to diffusion in the vapor,⁴ gaseous ions,¹² adsorption of impurities,⁴ chemical reaction,⁹ and dissipation of the heat of condensation have been omitted from the above discussion.

G. M. POUND

References

1. Band, W., "Quantum Statistics," New York, D. Van Nostrand, 1955.
- 1a. Hill, T. L., "Statistical Mechanics," New York, McGraw-Hill Book Co., 1956.
2. Volmer, M., "Kinetik der Phasenbildung," Dresden and Leipzig, Steinkopff, 1939.
3. Frenkel, J., "Kinetic Theory of Liquids," London, Oxford University Press, 1946.
4. Hirth, J. P., and Pound, G. M., "Condensation and Evaporation, Nucleation and Growth Kinetics," Oxford, Pergamon Press, 1963.
5. Wilson, C. T. R., *Phil. Trans. Roy. Soc. London*, **192**, 403; **193**, 289 (1899).

* Or edge dislocations with a component of the Burgers vector perpendicular to the surface.

6. Powell, C. F., *Proc. Roy. Soc. London*, **119**, 553 (1928).
7. Kossel, W., *Nachr. Akad. Wiss. Göttingen, Math. Physik Kl. I*, 135 (1927).
8. Stranski, I. N., *Z. Phys. Chem.*, **136**, 259 (1928).
9. Stranski, I. N., and Hirschwald, W., "Proceedings International Symposium on Condensation and Evaporation of Solids," U.S. Air Force, Dayton, 1962.
10. Burton, W. K., Cabrera, N., and Frank, F. C., *Phil. Trans. Roy. Soc. London Ser. A*, **243**, 299 (1950).
11. Dekeyser, W., and Amelinckx, S., "Les Dislocations et la Croissance des Cristaux," Paris, Masson 1956.
12. Thomson, J. J., "Conduction of Electricity through Gases," Cambridge, The University Press, 1906.

Cross-references: VAPOR PRESSURE AND EVAPORATION; STATES OF MATTER; CRYSTALLIZATION.

CONDUCTIVITY, ELECTRICAL

The electrical conductivity of a material is a measure of its ability to carry an electric current. The conductivity σ of an isotropic material in a steady (dc) electric field E is defined by

$$j = \sigma E$$

where j is the current density (charge transported per unit time across unit area perpendicular to the current flow). In mks units (used throughout this article) j is measured in amperes per square meter and E in volts per meter, so that σ is in $(\text{ohm-m})^{-1}$ or mhos per meter. The reciprocal of the conductivity is the electrical resistivity, $\rho = 1/\sigma$. The electrical resistance R of a sample is the ratio of the potential drop across the sample to the total current through the sample. The resistance of a cylindrical sample (with the current flow parallel to the axis of the cylinder) is

$$R = \rho l/A$$

where l is the length of the cylinder and A is its cross-sectional area. If the dimensions are measured in meters and ρ is in mks units, the resistance is given in ohms. The reciprocal of the resistance is the electrical conductance.

Many homogeneous solids and liquids obey Ohm's law for sufficiently small electric fields. Ohm's law states that the current through a sample is proportional to the potential drop across the sample, and thus R , ρ , and σ are independent of the impressed electric field. Samples which do not obey Ohm's law are called *non-linear*; gases fall into this category, as do many important circuit elements such as vacuum tubes and transistors. Most metals obey Ohm's law for fields up to at least 10^8 volts/m, though under certain conditions, semiconductors have shown deviations from Ohm's law for fields as low as 10^{-2} volts/m.

The existence of a finite resistance means that the energy delivered by the electric field to the current carriers is dissipated, being converted to

heat (mostly energy of atomic vibrations). The rate of dissipation per unit volume is given by

$$\frac{1}{2} \rho j^2 = \frac{1}{2} \sigma E^2$$

and is called the Joule heat.

In some solids, the conductivity is *anisotropic*, i.e., the magnitude of j depends upon the direction of E as well as its magnitude, and j and E are not necessarily parallel. If Ohm's law is obeyed it becomes

$$j = \sigma \cdot E$$

where σ is a second rank tensor. Anisotropic solids include single crystals of materials which do not have cubic crystal structures, and polycrystalline aggregates of such materials in which there exists some preferred orientation such as can be produced by extrusion or cold rolling.

If the applied electric field varies sinusoidally in time (ac), the conductivity generally depends upon the applied frequency ν . Appreciable deviations from the dc value may appear for microwave frequencies or greater. In the microwave and optical range, it is common to identify $\sigma = 2\pi\epsilon_0\nu$ as the imaginary part of the *dielectric coefficient*, where ϵ_0 is the permittivity of free space $1/4\pi\epsilon_0 = 9 \times 10^9$ newton m²/coulomb². Thus σ is closely related to the absorption of electromagnetic energy.

Solids are usually classified as metals, semiconductors or insulators. Metals are characterized by an increasing resistivity with increasing temperature. Resistivities of metals at room temperature range from about 10^{-8} to 10^{-6} ohm-m. Semiconductors are characterized by a decreasing resistivity with increasing temperature (impure semiconductors may show this behavior only at high temperatures). Resistivities of semiconductors at room temperature range from about 10^{-3} to 10^{-7} ohm-m. Insulators share the same temperature behavior of resistivity as semiconductors, so that the difference between the two classes of materials is one of degree only. Generally, materials with room temperature resistivities greater than 10^7 ohm-cm are called insulators. Resistivities as high as 10^{18} ohm-m have been observed. There are some materials intermediate between metals and semiconductors. For example, the resistivity of manufactured carbon decreases with increasing temperature at low temperatures, and increases at high temperatures. The room-temperature resistivity is also intermediate, varying from 10^{-5} to 10^{-4} ohm-m depending upon conditions of manufacture.

Except in the case of some insulators, the current in solids is carried by electrons. Quantum mechanics has shown that not all the electrons in a solid are free to carry current. The conductivity depends upon the number of free carriers and their ease of motion. The latter factor is expressed by the mobility μ which is defined by

$$v_d = \mu E$$

where v_d is the drift velocity (average velocity of the free carriers produced by the action of the

electric field). The drift velocity is usually very much smaller than a typical carrier velocity v (the average of the magnitude of the carrier velocities). The conductivity is then given by

$$\sigma = ne\mu$$

where n is the density of free carriers (in m⁻³), e is the magnitude of the electronic charge (1.6×10^{-19} coulomb), and μ is in square meters per volt per second. If more than one group of carriers is present, the total conductivity is the sum of contributions from each group. Quantum mechanics has also shown that the mobility in a pure, perfectly regular crystal would be unlimited. The mobility is limited by the scattering of the carriers by deviations such as the thermal vibrations of the atoms, impurity atoms, or irregularities in the crystal structure such as vacancies and dislocations. A simple approximate theory of the mobility allows it to be expressed as

$$\mu = e\tau/4\pi\epsilon_0 m^* \quad e\lambda/4\pi\epsilon_0 m^* v$$

where τ is the relaxation time (roughly, the average time between the collisions which a carrier makes with the scattering centers), λ is the mean free path of the carriers, and m^* is the effective mass of the carrier (a concept which comes from the theory of energy bands in solids, see SEMICONDUCTOR). If more than one scattering process is involved, the reciprocal of the relaxation time (scattering rate) is approximately the sum of contributions from each process. At room temperature, mobilities range from very small values (such as 10^{-1} m²/volt sec) to 10^2 m²/volt sec, relaxation times range from about 10^{-14} second to about 10^{-12} second, and mean free paths range from about 10^{-9} meter to about 10^{-6} meter. One of the early triumphs of quantum mechanics was the explanation of why the mean free path can be so much larger than the distance between neighboring atoms (of the order of 10^{-10} meter).

In good metals, n is the density of valence electrons, and is thus independent of temperature (except for very small temperature dependence due to the thermal expansion). Because the electrons follow the *Fermi-Dirac* distribution law, the typical velocity v is the velocity at the *Fermi Surface*, which is large (usually about 10^6 m/sec) and independent of temperature. The temperature dependence of the conductivity is that of the relaxation time τ , and the approximate additivity of scattering rates due to different processes results in *Matthiessen's rule* which states that the resistivity is approximately the sum of a temperature dependent part R_T , due to scattering by lattice vibrations, and a part R_i proportional to the concentration of impurities and lattice defects. As the amplitude of the lattice vibrations increases with increasing temperature, the scattering effect and R_T increase. Above the Debye temperature θ (given by $k\theta = h\nu_m$, where k is Boltzmann's constant and ν_m is the maximum frequency of the lattice vibrations), R_T is proportional to the absolute temperature. At low temperatures, R_T is proportional to the fifth power of the absolute

temperature, and at all temperatures it is well approximated by the *Bloch-Grüneisen formula*. At very low temperatures (1 to 18°K), some metals become superconductors and all measurable resistance disappears (see SUPERCONDUCTIVITY).

Pure semiconductors at absolute zero temperature have no free electrons. As the temperature is increased, some electrons are excited to current-carrying states (in the *conduction band*). The states that are left unoccupied (in the *valence band*) are also free to carry current, and are called *free holes*. The concentrations of electrons and holes increase very rapidly with temperatures, causing the resistivity to decrease. The carrier concentration may be expressed as

$$n = f(T) \exp(-\Delta E/2kT)$$

where T is the absolute temperature, ΔE is the energy gap between the valence and conduction bands, and $f(T)$ is a slowly varying function of temperature. Extra carriers may also be provided by impurity atoms, *donors* contributing electrons and *acceptors* trapping electrons and producing holes. Very small impurity concentrations may make very large changes in resistivity. If most of the free carriers come from impurities, the semiconductor is called *extrinsic*; it is *n*(negative)-type if electrons predominate or *p*(positive)-type if holes predominate. If most of the carriers come from thermal excitation, the concentration of electrons and holes are about equal, and the material is called *intrinsic*. The rate of change with temperature of the mobility of a semiconductor is generally less than the rate of change of carrier concentration. The mobility is often represented by $\mu = cT^r$, where r varies from about -2.2 to +1.5 depending upon the particular material, concentration, and type of defects, etc. The conductivity of intrinsic material can also be expressed as

$$\sigma = g(T) \exp(-\Delta E/2kT)$$

where $g(T)$ is another slowly varying function of T . This equation is often used to analyze experimental data and find ΔE .

Insulators may be thought of as semiconductors with such large energy gaps that n is very small and the resistivity is very high. In addition, the ions in some insulating solids (such as the alkali halides) are free to move and carry a measurable current. The ions move by "hopping" into vacant lattice sites. As they must cross a potential energy barrier, the ionic mobility is proportional to an activation factor $\exp(-\epsilon/kT)$. The resistivities of such solids are of the order of 10^2 to 10^8 ohm-m at room temperature, but are as low as 10^{-3} to 1 ohm-m at elevated temperatures.

Another mechanism for electronic conductivity seems to operate in certain oxides and perhaps in some organic semiconductors. In this case, the electrons are localized and move by "hopping" in the same way as the ions in ionic solids. This phenomenon is currently being investigated.

Because the resistivity depends upon the state

of crystalline order, it is used as a tool in studying phase changes, such solid-liquid, order-disorder, magnetic, and crystal structure transitions. The resistivity of some materials is affected by pressure and mechanical strain. Many insulators and semiconductors exhibit the phenomenon of PHOTOCONDUCTIVITY, in which the absorption of light produces free carriers and increases the conductivity.

Many conductors show the phenomenon of *magnetoresistance*, which is the increase in resistance when the conductor is placed in a magnetic field. (A very few materials exhibit negative magnetoresistance, which seems to be related to inhomogeneous structure of the conductor.) The basic cause of the magnetoresistance is the Lorentz force, which causes the electrons to move in curved paths between collisions. Even for isotropic materials, the conductivity and resistivity must be taken as tensors in the presence of the magnetic field, called the *magnetoconductivity tensor* and the *magnetoresistivity tensor*. The off-diagonal elements of the magnetoresistivity tensor are related to the *Hall effect*. The on-diagonal elements for isotropic materials or for the magnetic field parallel to a crystal axis are usually called the longitudinal magnetoresistance (for the current parallel to the magnetic field) and transverse magnetoresistance (for the current perpendicular to the magnetic field). For small values of the magnetic field, the change in resistance is proportional to the square of the magnetic field strength. For large magnetic fields, the resistance may saturate (approach a constant value), continue to increase as the square of the magnetic field strength, or follow a more complicated behavior. The magnetoresistance ratio (change in resistance divided by the zero-field resistance) reaches only a few per cent for most materials, but at low temperatures for some materials, such as bismuth, approaches 10^6 . In general, the transverse effect is larger than the longitudinal one; the effect is larger in high-mobility materials and is largest for materials with more than one type of charge carrier. The magnetoresistance may be used in conjunction with the Hall effect to deduce the type, density, and mobility of charge carriers, and to obtain information about the FERMI SURFACE.

J. W. MCCLURE

References

- Pugh, E. M., and Pugh, E. W., "Principles of Electricity and Magnetism," Ch. 6, Reading, Mass., Addison-Wesley Publishing Co., 1960. Gives a more complete discussion of electrical conductivity.
- Dekker, A. J., "Solid State Physics," Chs. 11, 12 and 13, Englewood Cliffs, N.J., Prentice-Hall, 1957. Introductory treatment of electrical conductivity.
- Ehrenberg, W., "Electric Conduction in Semiconductors and Metals," London, Oxford University Press, 1958. Advanced treatment of electrical conductivity.

CONSERVATION LAWS AND SYMMETRY

Among the most basic of the laws of nature are the conservation laws. A conservation law is a statement which says that in a given physical system under specified conditions, there is a certain measurable quantity which never changes regardless of the actions which go on within the system. One of the tasks of physics is to determine what those special quantities are which do not change under a given set of conditions.

Three classical conservation laws are those of energy, momentum and angular momentum. Laws of this nature are postulated as a result of many measurements of the energies and momenta involved in reactions of all kinds. It is always found that within the limit of accuracy of the experiment, the amount of energy in the system after the reaction is the same as the amount of energy before the reaction. Prior to the development of elementary particle physics, there was much emphasis on transformation of energy between its various "forms," such as mechanical, electrical, thermal, etc. The modern viewpoint is that the macroscopic behavior of matter is the result of interactions between elementary particles and that these elementary interactions individually obey the various conservation laws. In its elementary form, the law of conservation of energy states that when two or more particles interact, the total energy (kinetic plus potential) is always a constant. When we say, for example, that the kinetic energy of a moving object has been transformed into the potential energy of a compressed spring, what we mean on a microscopic level is that the atoms of the spring have been pressed closer together so that there is a greater amount of potential energy in the electric fields between the atoms of the spring. This increase of potential energy is associated with an equal decrease in the kinetic energy of the object which caused the spring to compress.

An important function of the conservation laws is that they allow us to make many predictions about the behavior of a system without going into the mechanical details of what happens during the course of a reaction. They give us a direct connection between the state of the system before the reaction and the state after the reaction. In particular we can say that any action which violates one of the conservations laws must be forbidden.

With the development of the Lagrangian and Hamiltonian methods of solving physical problems, and particularly with the growth of importance of quantum mechanics, it has become clear that the conservation laws are closely connected with the concept of symmetry in space and time. This is based upon the fact that the interaction between two or more objects can be described in terms of a potential energy function. If the potential energy of the system is known for any position of these objects in space and time, then we can predict the future motions of the objects. However, certain predictions can be

made without going through a complete solution of the equations of motion. For example, it is found that if the potential energy does not depend explicitly on one of the space coordinates, then the momentum associated with that coordinate never changes—it is a constant of the motion.

The following examples illustrate the major types of "geometrical symmetries" encountered in classical physics.

(1) An object moves in a three-dimensional space where its potential energy is the same at every point. The expression describing the potential does not explicitly contain the coordinates x , y or z . That is, the system is invariant with respect to translation of the origin of the coordinate system in any direction. This symmetry is associated with conservation of linear momentum: the momentum in all three dimensions is a constant.

(2) An object moves in a world which is flat so that the force of gravity is in the vertical (z) direction. The potential depends only on the height of the object above the ground, but does not depend on its location in the horizontal plane; i.e., the potential is invariant to a translation of the coordinate system in the x - y plane. There is now symmetry in two dimensions, and momentum in the x - y plane is conserved.

(3) In the interaction between two spherical bodies, if the potential depends only on the distance between the bodies, then there is spherical symmetry. The system is invariant to a rotation of the coordinate system about any axis. In this case, two components of angular momentum are conserved: as the two bodies orbit around their common center of mass, the magnitude of their angular momentum is constant, while the plane of the orbit in space never changes.

(4) If the interaction between two objects does not depend explicitly on the time coordinate, then the actions which take place do not depend on when we start measuring time; i.e., the properties of the system are invariant with respect to a translation of the origin of coordinates along the time axis. This symmetry is associated with conservation of energy. Use of a four-dimensional coordinate system allows us to associate conservation of momentum and energy in a unified manner with the geometrical symmetry of space-time.

(5) Another conservation law known in classical physics is that of conservation of electric charge. Since electric charge comes in discrete quantities (there is no known way of breaking an electric charge down into bundles smaller than that contained in an electron), this law deals with the counting of objects, rather than with the measurement of continuous variables, such as momentum or energy. Conservation of electric charge means that the total number of electric charges (taking positive and negative signs into account) in a closed system never changes. This formerly meant that electric charge could not be created or destroyed. We now speak of the creation of charged particles, but the creation

of a positive charge must always be accompanied by formation of an equal negative charge (e.g., an electron-positron pair is created by a photon). Conservation of electric charge is associated with a symmetry property of Maxwell's equations known as gauge invariance, which states that the absolute value of the electrical potential (as opposed to the relative value) plays no part in physical processes. More recently, in the development of quantum mechanics, it appears that conservation of electric charge is connected with the fact that the properties of a system of particles do not depend on the phase of the wave function describing the system.

The conservation laws and the concept of symmetry have acquired new importance in the area of elementary particle physics. The conservation laws act as "selection rules" to determine which reactions may take place between the many existing particles out of the enormous number of otherwise conceivable reactions.

An important property of elementary particles is *parity*. Each particle has a parity number associated with it: either $+1$ or -1 , depending on the type of particle. In an assembly of particles, there is a total parity, which is the sum of the individual parities. If parity is conserved, then this total parity does not change during the course of the reaction. This property of matter is associated with the so-called mirror symmetry. All the laws of nature which possess this type of symmetry are such that if the words "right" and "left" are interchanged in the statement of the law, then the behavior of the system obeying these laws is unchanged. It was formerly believed that every natural law was of this type. As a result of this conservation of parity, it was believed to be impossible to describe the difference between "right" and "left" by the use of words alone. A famous paper by C. N. Yang and T. D. Lee in 1956 pointed out the fact that in a special class of reactions involving the "weak nuclear" interaction, parity need not be conserved. As a result of this, it was seen that the universe does possess an asymmetry between right and left and that it is possible to describe an experiment which will definitely distinguish between the directions "right" and "left" in this universe.

Another type of symmetry of importance in elementary particle physics is that entitled *charge conjugation*. This principle states that if each particle in a given isolated system is replaced by its corresponding antiparticle, then nobody will be able to tell the difference. For example, if in a hydrogen atom the proton is replaced by an anti-proton and the electron is replaced by a positron, then this antimatter atom will behave exactly like an ordinary atom, as long as it does not come into contact with ordinary atoms.

However, it turns out that there are certain types of reactions where this rule does not hold, and these are just the types of reactions where conservation of parity breaks down. For example, consider a piece of radioactive material emitting electrons by beta decay. The radioactive nuclei are lined up in a magnetic field which is produced

by electrons traveling clockwise in a coil of wire, as seen by an observer looking down on the coil. Because of the asymmetry of the radioactive nuclei, most of the emitted electrons travel in the downward direction. If the same experiment were done with similar nuclei composed of antiparticles and the current in the magnet coil consisted of positrons instead of electrons, then the emitted positrons would be found to travel in the upward, rather than in the downward, direction. Interchanging each particle with an antiparticle has produced a change in the experiment.

However, the symmetry of the situation can be restored if we interchange the words "right" and "left" in the description of the experiment at the same time that we exchange each particle with its antiparticle. In the above experiment, this is equivalent to replacing the word "clockwise" by "counterclockwise." When this is done, the positrons are emitted in the downward direction, just as the electrons in the original experiment. The laws of nature are thus found to be invariant to the simultaneous application of charge conjugation and mirror inversion.

Other more technical conservation laws play a role in elementary particle physics. *Conservation of baryon number* and *conservation of strangeness* are rules required to account for the fact that certain reactions involving heavy particles are forbidden. *Time reversal invariance* describes the fact that in reactions between elementary particles, it does not make any difference if the direction of the time coordinate is reversed.

It is a temptation to say that "nature likes symmetries" in order to prove the necessity of a conservation law. However, to avoid pitfalls it must be realized that it is human beings who like symmetries. While a symmetry idea may first suggest a conservation law, the conservation law must be tested by experiment to see if the symmetry is valid.

MILTON A. ROTHMAN

References

- Weyl, H., "Symmetry," Princeton, N.J., Princeton University Press, 1952.
- Yang, C. N., "Elementary Particles," Princeton, N.J., Princeton University Press, 1962.
- Rothman, M. A., "The Laws of Physics," New York, Basic Books, Inc., 1963.
- Feynman, R. P., Leighton, R. B., and Sands, M., "The Feynman Lectures on Physics," Reading, Mass. Addison-Wesley Publishing Co., 1963.
- Chew, G. F., Gell-Mann, M., and Rosenfeld, A. H., "Strongly Interacting Particles," *Sci. Am.* (February, 1964).
- Morrison, P., "The Approximate Nature of Physical Symmetries," *Am. J. Phys.* (September, 1958).

Cross-references: ANTIPARTICLES; ELEMENTARY PARTICLES; IMPULSE AND MOMENTUM; POTENTIAL; ROTATION—CIRCULAR MOTION; WEAK INTERACTIONS; WORK, POWER AND ENERGY.

CONSTANTS, FUNDAMENTAL

It is basic to the structure of modern physical theory that physical quantities such as the rest mass of the electron or the speed of light in vacuum have fixed and unchanging numerical values. It is also true that our knowledge of the numerical magnitude of these values is variable and, in general, changes with each new measurement which is made. For example, measurements of the velocity of light made over the past century have yielded results which have varied over wide ranges. This has even led some theorists to suggest that the velocity of light may in fact not be an exact constant but might have periodic variations with periods of several years. Not only does such a suggestion violate established theories of relativity, but it also violates the experimental evidence of a host of well-established indirect experiments. As a result of these indirect experiments, one can only conclude that it must be the experimental technique which is variable, rather than the speed of electromagnetic radiation. The uniqueness of the magnitudes of such quantities as the charge and mass of the electron or of Planck's constant of action has been well established by various indirect measurements and is fundamental to the current structure of physical theory. One can say that the existence of unique magnitudes is a characteristic of the fundamental nature of these concepts.

Fundamental Units. The numerical values of the fundamental constants are closely related to the question of units and standards, and it has long been suggested that fundamental physical constants be used as standards for physical units. It has been recognized that a fundamental physical quantity as a unit of length, for example, would be conceptually much better than the arbitrary unit of the foot or the meter. It is possible to consider as an appropriate unit of length the Bohr radius, $\hbar^2/4\pi^2me^2$, or alternately the Compton wavelength of the electron, h/mc . Because of the limitations of physical measurement, the use of such fundamental standards of length would be less accurate than other possible standards. However, the use of a physical unit rather than a completely arbitrary one (such as a standard meter bar) has the advantage of unique reproducibility and universal availability, and the unit of length is defined in terms of the wavelength of the electromagnetic radiation arising from the optical transition $2p_{10} - 5d_5$ in the isotope of krypton of mass 86. In terms of the wavelength of this radiation, the meter is now defined as 1650763.73 wavelengths, and the foot is defined by International agreement among all of the English-speaking nations to be exactly 0.3048 meters.* The unit of time (the second) is defined in terms of the tropical year, although the definition of the second in terms of the frequency of some atomic standard is being seriously considered and one strong front-runner in this race is the frequency of the hyperfine splitting in

hydrogen gas as maintained in the hydrogen maser. This frequency is perhaps one of the most accurate physical measurements known today $\nu = 1420405751.800 \pm 0.028 \text{ sec}^{-1}$ †

Experimental Determinations. The experimental determination of the numerical values of the physical constants is uncertain and variable if we confine our attention to the limits of experimental capabilities. The difficulties inherent in measuring physical constants to an accuracy of a few parts per million are great. Direct measurements of the mass of the electron or of the electronic charge are not as accurate as measurements which determine instead various combinations of these quantities. Whereas Millikan was able to measure the elementary charge on an electron in 1912, to one part in a few thousand (and this experiment can hardly be improved upon today), our current knowledge of the electronic charge, with an accuracy of approximately 1 part in 100,000 comes from combining measurements of the Sommerfeld fine structure constant with the gyromagnetic ratio of the proton, the Faraday constant, the magnetic moment of the proton relative to the Bohr magneton, and a half-dozen or so other measurements which affect the final result to a greater or lesser degree. In fact, the present knowledge of the numerical values of all of the so-called fundamental constants of physics come from such indirect measurements. Of these physical constants, only the universal gravitational constant, G , is measured independently of the others. It is true, in the first place, that no theoretical relation is known which relates G to the physical constants and, in the second place, that the accuracy with which G is known is several orders of magnitude poorer than the accuracy of the atomic constants.

With the growth in our knowledge of natural laws and of the technical means of making precise physical and chemical measurements, an increasing number of relationships have been discovered between the fundamental constants of physics and chemistry. At the present time, the situation regarding our knowledge of these constants can be considered as resembling a multidimensional bridge truss made up of elastic members, in which the length of each member represents the experimentally measured relationship between constants, and the stiffness is a measure of the

† At its meeting of October 6, 1964, the Twelfth General Conference on Weights and Measures adopted the following definition of the second of time in terms of the "atomic clock", superseding the previous astronomical definition in terms of the length of the mean solar year. "The standard to be employed is the transition between the two hyperfine levels $F = 4, M_F = 0$ and $F = 3, M_F = 0$ of the fundamental state $^2S_{1/2}$ of the atom of cesium 133 undisturbed by external fields and the value $9\,192\,631\,770$ hertz [cycles per second] is assigned." To within the accuracy of the respective measurements, this does not entail any change in the size of the second, but does provide a more accurate and more easily usable standard by which to measure the length of a second.

* Thus the English-speaking world uses the metric system for standards, if not for units.

accuracy of this measurement. The problem is to determine the positions of the nodes of this network. Furthermore, the alteration of any one member will produce an effect which will be transmitted throughout the entire structure. This will be true whether we change the length of a member, or its stiffness, or remove it entirely.

In order to determine the values of the physical constants from such an overdetermined set of data, it has become common to use the method of least squares. This can be considered as equivalent to the problem of minimizing the stored potential energy in our multidimensional framework. The fundamental requirement, however, is basically one of establishing a method of analysis which is consistent and independent of the choice of variables used to describe the situation. The method of least squares not only does this but also provides a procedure which yields "best" values of the constants in the sense that these values are the most accurate. For these reasons, least squares adjustment has been used for all of the significant determinations of the values of the fundamental physical constants over the past 15 years. The unit of mass and the unit of temperature, however, are still defined in a completely arbitrary manner—the unit of mass, in terms of the mass of the prototype kilogram, and the unit of temperature, as $1/273.16$ of the absolute thermodynamic temperature of the triple point of pure water.

Numerical Values. Because the numerical values are the result of a least squares adjustment, care must be used in computing the standard deviation of quantities whose values are not given in a table. Because of the interdependence of the numerical data, the errors are strongly correlated. The errors in these constants may be either greater or less than those computed by simple propagation of error, depending on whether the correlation is positive or negative. Thus the relative uncertainty in a computed constant not given in a table can be obtained only from a calculation which includes the full variance matrix of the least squares adjustment. For more

complete details on the point, see "Present Status of our Knowledge of the Numerical Values of the Fundamental Physical Constants," E. Richard Cohen and J. W. M. DuMond, Second International Conference on Nuclidic Masses, Vienna, Austria, Springer-Verlag, 1964.

Tables 1, 2 and 3 present a partial listing of the current best values of the fundamental physical

TABLE 1. DEFINED VALUES AND EQUIVALENTS

Meter (m)	1650763.73 wavelengths of the transition $2p_{10} - 5d_5$ in ^{86}Kr
Kilogram (kg)	Mass of the international kilogram
Second (sec)	$1/31556925.9747$ of the tropical year at 12 hours ET, 0 January, 1900 (yr = 365 days, 5 hours, 48 minutes, 45.9747 seconds)
Degree Kelvin (°K)	In the thermodynamic scale, $273.16^\circ\text{K} =$ triple point of water $T(^{\circ}\text{C}) = T(^{\circ}\text{K}) - 273.15$ (freezing point of water, $0.0000 \pm 0.0002^\circ\text{C}$)
Unified atomic mass unit (<i>u</i>)	$1/12$ the mass of an atom of the ^{12}C nuclide
Standard acceleration of free fall (<i>g_n</i>)	9.80665 m/sec^2 980.665 cm/sec^2
Normal atmosphere (atm)	101325 N/m^2 $1013250 \text{ dyne/cm}^2$
Thermochemical calorie (cal _{th})	4.184 J
International Steam Table calorie (cal _{IT})	4.1868 J $4.1868 \times 10^7 \text{ erg}$
Liter	0.00100028 m^3 1000.028 cm^3 (recommended by CIPM, 1950)
Inch (in.)	0.0254 m 2.54 cm
Pound (adp) (lb)	0.45359237 kg 453.59237 g

TABLE 2. GENERAL PHYSICAL CONSTANTS

Based on the analysis of E. R. Cohen and J. W. M. DuMond, "Proceedings of the Second International Conference on Nuclidic Masses and Related Constants," Springer-Verlag, Vienna, Austria 1964. These values have been recommended for general use by the National Academy of Sciences—National Research Council and have been adopted by the U.S. National Bureau of Standards *Natl. Bur. Std. Tech. News Bull.* 175 (October, 1963).

Digits in parentheses represent 3 standard deviation error limits in the final digits of the quoted value.

C - coulomb G - gauss Hz hertz J joule N - newton
T - tesla u - unified nuclidic mass unit W - watt Wb - weber

Constant	Symbol	Value	UNIT	
			mksa	cgs
Speed of light in vacuum	<i>c</i>	2.997925(3)	10^8 m/sec	$\times 10^{10} \text{ cm/sec}$
Gravitational constant	<i>G</i>	6.670(15)	$10^{-11} \text{ Nm}^2 \text{ kg}^{-2}$	$10^{-8} \text{ dyn cm}^2 \text{ g}^{-2}$
Elementary charge	<i>e</i>	1.60210(7)	10^{-19} C	$10^{-20} \text{ cm}^{1/2} \text{ g}^{1/2} \text{ sec}^{-1/2}$
		4.80298(20)		$10^{-10} \text{ cm}^{3/2} \text{ g}^{1/2} \text{ sec}^{-1/2}$
				$10^{-10} \text{ cm}^{3/2} \text{ g}^{1/2} \text{ sec}^{-1/2}$
Avogadro constant	<i>N_A</i>	6.02252(28)	$10^{20} \text{ kmole}^{-1}$	$10^{23} \text{ mole}^{-1}$
Mass unit	<i>u</i>	1.66043(7)	10^{-27} kg	10^{-24} g
Electron rest mass	<i>m_e</i>	9.1091(4)	10^{-31} kg	10^{-28} g

TABLE 2—Continued.

Constant	Symbol	Value	UNIT	
			mksa	cgs
Proton rest mass	m_p	5.48597(9) 1.67252(8) 1.00727663(24)	10^{-4} u 10^{-27} kg u	10^{-4} u 10^{-24} g u
Neutron rest mass	m_n	1.67482(8) 1.0086654(13)	10^{-27} kg u	10^{-24} g u
Faraday constant	F	9.64870(16) 2.89261(5)	10^4 C/mole	10^3 cm ^{1/2} g mole ⁻¹ * 10^{14} cm ^{3/2} g ^{1/2} sec ⁻¹ mole ⁻¹ **
Planck constant	h	6.6256(5)	10^{-34} J sec	10^{-27} erg sec
Fine structure constant, $2\pi e^2/hc$	$h/2\pi$	1.05450(7)	10^{-34} J sec	10^{-27} erg sec
α	α	7.29720(10)	10^{-3}	10^{-3}
Charge to mass ratio for electron	$1/\alpha$	137.0388(19)		
Quantum of magnetic flux	e/m_e	1.758796(19) 5.27274(6)	10^{11} C/kg	10^7 cm ^{1/2} g ^{-1/2} * 10^{17} cm ^{3/2} g ^{-1/2} sec ⁻¹ **
Rydberg constant	hc/e	4.13556(12)	10^{-11} Wb	10^{-7} G cm ²
Bohr radius	h/e	1.37947(4)		10^{-17} cm ^{1/2} g ^{1/2} **
Compton wavelength of electron	R_∞	1.0973731(3)	10^7 m ⁻¹	10^5 cm ⁻¹
Electron radius $e^2/m_e c^2 = r_e$	a_0	5.29167(7)	10^{-11} m	10^{-9} cm
Thomson cross section	$h/m_e c$	2.42621(6)	10^{-12} m	10^{-10} cm
Compton wavelength of proton	$\lambda_C/2\pi$	3.86144(9)	10^{-13} m	10^{-11} cm
Gyromagnetic ratio of proton	$\lambda_C/2\pi$	2.81777(11)	10^{-15} m	10^{-13} cm
(uncorrected for diamagnetism, H ₂ O)	$8\pi r_e^2/3$	6.6516(5)	10^{-29} m ²	10^{-25} cm ²
Bohr magneton	$\lambda_{C,p}/2\pi$	1.32140(4)	10^{-15} m	10^{-13} cm
Nuclear magneton	γ	2.67519(2)	10^8 rad sec ⁻¹ T ⁻¹	10^4 rad sec ⁻¹ G ⁻¹ *
Proton moment	$\gamma/2\pi$	4.25770(3)	10^7 Hz/T	10^3 sec ⁻¹ G ⁻¹ *
(uncorrected for diamagnetism, H ₂ O)	γ'	2.67512(2)	10^8 sec ⁻¹ T ⁻¹	10^4 rad sec ⁻¹ G ⁻¹ *
Gas constant	$\gamma'/2\pi$	4.25759(3)	10^7 Hz/T	10^3 sec ⁻¹ G ⁻¹ *
Boltzmann constant	μ_B	9.2732(6)	10^{-24} J/T	10^{-21} erg/G*
First radiation constant ($2\pi hc^2$)	μ_N	5.0505(4)	10^{-27} J/T	10^{-24} sec/G*
Second radiation constant	μ_p	1.41049(13)	10^{-26} J/T	10^{-23} erg/G*
Stefan-Boltzmann constant	μ_p/μ_N	2.79276(7) 2.79268(7)	.	.
Gas constant	R	8.3143(12)	J deg ⁻¹ mole ⁻¹	10^7 erg deg ⁻¹ mole ⁻¹
Boltzmann constant	k	1.38054(18)	10^{-23} J/deg	10^{-16} erg/deg
First radiation constant ($2\pi hc^2$)	c_1	3.7415(3)	10^{-16} W m ²	10^{-5} erg cm ² sec ⁻¹
Second radiation constant	c_2	1.43879(19)	10^{-2} m deg	cm deg
Stefan-Boltzmann constant	σ	5.6697(29)	10^{-8} W m ⁻² deg ⁻⁴	10^{-5} erg cm ⁻² sec ⁻¹ deg ⁻⁴

* Electromagnetic units.

** Electrostatic units.

constants. Most of these values have been recommended for general use by the U. S. National Academy of Sciences, National Research Council Committee on Fundamental Physical Constants (1963), and have been adopted by the U. S. National Bureau of Standards.

In these Tables, the figures in parentheses represent limits of error taken to be three times the standard deviation calculated in the least squares adjustment. If it were safe to assume that the error distributions are gaussian, these error limits would represent a range such that there was

a probability of only 1 chance in 400 that the correct value would lie outside those limits. The use of three standard deviations is justified in this case, however, because of uncertainties regarding systematic errors and/or theoretical inadequacies in the interpretation of the experimental data. All of the data in these Tables which depend upon molecular weights are expressed on the unified scale adopted in 1960 by the International Union of Pure and Applied Physics and the International Union of Pure and Applied Chemistry. This scale was based on the arbitrary assignment of the

TABLE 3. ENERGY CONVERSION FACTORS

1 eV	= $1.60210(7) \times 10^{-19}$ J
	= $1.60210(7) \times 10^{-12}$ erg
	= $8065.73(23)$ cm ⁻¹
	= $2.41804(7) \times 10^{14}$ sec ⁻¹
Vλ	= $12398.1(0.4) \times 10^{-8}$ eV cm
1 eV/particle	= $11604.9(1.5)$ °K
	= $23060.9(4)$ cal _{th} mole ⁻¹
	= $23045.5(4)$ cal _{IT} mole ⁻¹
1 amu	= $931.478(15)$ MeV
Proton mass	= $938.256(15)$ MeV
Neutron mass	= $939.550(15)$ MeV
Electron mass	= $511006(5)$ eV
Rydberg constant	= $13.6054(4)$ eV
Gas constant, R_0	= 8.31433×10^7 erg mole ⁻¹ deg ⁻¹
	= 0.0820538 atm mole ⁻¹ deg ⁻¹
	= 82.0561 cm ³ atm mole ⁻¹ deg ⁻¹
	= 1.98717 cal _{th} mole ⁻¹ deg ⁻¹
	= 1.98584 cal _{IT} mole ⁻¹ deg ⁻¹
Standard volume of an ideal gas, V_0	22413.6 cm ³ mole ⁻¹

mass of exactly 12 units to the isotope ¹²C. As such, this definition replaces both the physical scale of atomic weights based on the assignment of mass 16 to the isotope ¹⁶O and the chemical scale of atomic weights which assigns the mass 16 to the "natural" isotopic mixture of oxygen isotopes.

E. RICHARD COHEN

CORIOLIS EFFECT

A marksman fires his rifle due north. In the absence of wind, he might well expect it to travel in a straight line and land due north of him. But will it? The physicist would, in general, answer no on the basis that the earth is rotating and is not, therefore, an inertial frame of reference (see ROTATION—CIRCULAR MOTION). G. G. Coriolis first analyzed this effect in 1844, and he is acknowledged in its name. For large artillery projectiles this effect may be significant, but for hand carried weapons it is usually negligible. For instance, a typical .22-caliber rifle bullet might be horizontally deflected one foot in traveling a mile.

That a projectile will normally follow a path which is curved in the horizontal leads conversely to the idea that to follow a straight path over the rotating earth requires the application of a side-wise force. Even though this appears to violate Newton's first law (see DYNAMICS), it is perfectly true in a non-inertial system—hence, the reason that this force is sometimes called "fictitious."

A body which is moving with constant speed in a straight line in an inertial system is not accelerating and is subject to no net force. An observer in a rotating coordinate system will, however, observe the same object to follow a curved path. The observer may treat this apparent deflection from a straight line as an acceleration which is always perpendicular to the path of the object. It is called the Coriolis acceleration. The

apparent force applied to the body to cause the deflection from a straight path into a curve is called the Coriolis force.

The true acceleration of a body moving with constant speed in a straight line in a rotating coordinate system is equal in magnitude to the apparent acceleration just described. An expression for it may be derived simply and quite rigorously, but not generally, for the case of a body moving with constant radial speed in a rotating system. It consists of two distinct components both easily evaluated for the case mentioned. One component arises from the change in direction of the radial velocity of the body, the other from its change in tangential velocity due to changing distance from the center of rotation (see ROTATION—CIRCULAR MOTION).

Suppose the rotating system has a constant angular velocity ω , and the body moves radially with constant speed v . It is initially at distance r_1 from the center of rotation with velocity \vec{v}_1 . After a small time interval Δt , it is at distance r_2 with velocity \vec{v}_2 (see Fig. 1). The two velocities \vec{v}_1

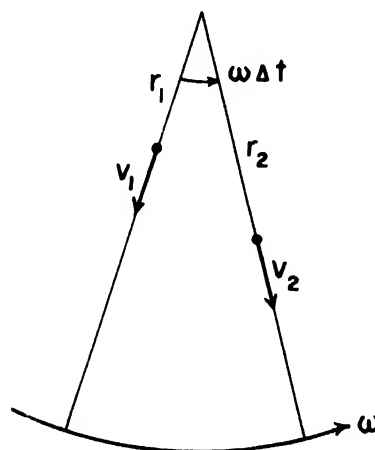


FIG. 1.

and \vec{v}_2 are related since the final velocity \vec{v}_2 is the vector sum of the initial velocity \vec{v}_1 and the change in velocity $\Delta \vec{v}$. Refer to Fig. 2 for a vector diagram of the preceding statement (see STATICS). If we consider instantaneous values, Δt approaches zero, the angle $\omega \Delta t$ approaches zero, and the chord Δv approaches its arc in length. We may then use the well known angle-arc relationship (arc length) = (radius) (angle in radians), and write

$$\Delta v = r \omega \Delta t, \text{ or } \Delta v / \Delta t = \omega v$$

Now $\Delta v / \Delta t$ is the acceleration component (a_1) resulting from the change in direction of \vec{v} .

The tangential velocity of a body equals the product of its angular velocity and its radius of rotation. Equating changes in these quantities yields

$$\Delta v_t = \omega(r_2 - r_1) = \omega \Delta r$$

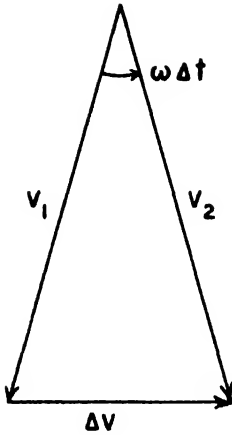


FIG. 2.

Or, dividing by Δt

$$\Delta v_1 / \Delta t \rightarrow \omega \Delta r / \Delta t$$

Now $\Delta v_1 / \Delta t$ is the acceleration component (a_2) arising from changing tangential velocity, and $\Delta r / \Delta t$ is the radial speed v . Hence a_2 also equals ωv .

Since both a_1 and a_2 lie in the same direction, being perpendicular to the radial velocity and to the right in the figure, their magnitudes may be added together with the sum equaling the magnitude of the Coriolis acceleration a_c

$$a_1 + a_2 = a_c = 2\omega v$$

Using Newton's second law $F = ma$, the Coriolis force is

$$F_c = 2m\omega v$$

The earth is not, of course, a rotating plane. The Coriolis acceleration reaches a maximum at the poles and vanishes at the equator. If v_h denotes horizontal velocity on the earth's surface, the resulting Coriolis acceleration may be evaluated by reference to Fig. 3 where ϕ is the latitude and v is the true radial speed with respect to the earth's axis.

$$v/v_h = \sin \phi$$

$$v = v_h \sin \phi$$

and the Coriolis acceleration becomes

$$a_c = 2\omega v_h \sin \phi$$

and the Coriolis force is

$$F_c = 2m\omega v_h \sin \phi$$

The Coriolis effect applies to any object moving on the surface of the earth, and a more general treatment will show it to be completely independent of the direction of motion. The quantity $2\omega \sin \phi$ is commonly known as the Coriolis parameter. Since the earth rotates 2π radians in 24 hours, or at a rate of 7.27×10^{-5} radians/sec, the parameter is quite small, and equals exactly 10^{-4} sec^{-1} at about $43\frac{1}{2}^\circ$ latitude. This small value

for the Coriolis parameter means that in everyday life its effects are small and go largely unnoticed. For instance, the Coriolis force on an automobile driving at turnpike speeds might typically be one pound. The acceleration can cause considerable deflection of long range artillery, however, and appropriate corrections must be made.

The Coriolis effect plays a large role in the great mass movements of the oceans and atmosphere. In the northern hemisphere, the apparent deflection is always to the right of the direction of motion. For example, air drawn toward a center of low pressure is deflected to the right and eventually flows around the low pressure area in a counterclockwise motion. This motion characterizes frontal storms typical of temperate climates. If the pressure force and Coriolis force are equal, the resulting wind velocity is said to be geostrophic (i.e., "turned by the earth"). The wind velocity in large storms above several thousand feet elevation is closely geostrophic.

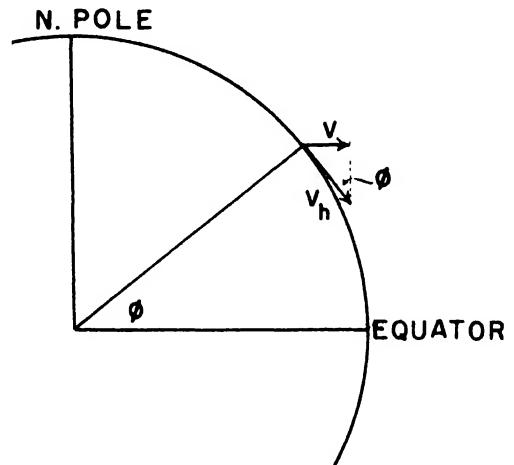


FIG. 3.

Though it has been only a little over a century since the first analysis of this effect, the Coriolis force has through the centuries influenced man's environment through its control of the motions of winds and waters, and hence the distribution of the sun's heat over the earth. Ancient man, not aware of the motion of his planet, was nevertheless profoundly influenced by it.

JULIAN M. PIKE

References

- Byers, Horace Robert, "General Meteorology," third edition, New York, McGraw-Hill Book Co., 1959.
- Coriolis, G. G., "Traite de la Mecanique de Corps Solides," Paris, 1844.
- Fowles, Grant R., "Analytical Mechanics," New York, Holt, Rinehart and Winston, 1962.
- Stephenson, Reginald J., "Mechanics and Properties of Matter," second edition, New York, John Wiley & Sons, Inc., 1960.

COSMIC RAYS

Cosmic rays are conventionally divided into two classes: primary and secondary. The former are, for the most part, energetic charged particles of extraterrestrial origin, while the latter are the products resulting from collisions of the primary cosmic rays with atoms of the earth's atmosphere.

The Primary Radiation. The primary cosmic radiation is quite striking in four respects: the primary intensity is essentially constant in time, isotropic in space, anomalous in composition, and it contains very energetic particles indeed.

Present-day measurements indicate that with the exceptions of the local perturbations discussed under solar effects, the primary cosmic-ray intensity exhibits less than 1 per cent variation in time. In fact, measurements of radioactivity produced in meteors by cosmic rays show that the intensity has not appreciably changed in the last several million years.

Isotropy simply means that there is no direction or directions in the sky from which the bulk of cosmic radiation emanates; this includes the direction of the sun. Since there are magnetic fields of varying magnitude and direction throughout interstellar space which deflect charged particles, the isotropic nature of the primary intensity is not too surprising. For one can say that, owing to "collisions" of the cosmic rays with these fields, the primary particles lose their "memory" of the directions to their source. Magnetic fields also act in other ways. For example, the earth's field effectively prevents particles of less than 10^8 eV energy from reaching the earth at all (1 eV 1.6×10^{-19} joule). Moreover, there is some indication that cosmic rays between 10^{14} and 10^{18} eV may be very slightly guided along the field lines of the local spiral arm of our galaxy. It should be remarked that the diameter of the orbit of a proton of 10^{18} eV moving in the galactic magnetic field would be comparable with the thickness of the galactic disc, so that above 10^{18} eV such guidance is impossible.

The total number of primary cosmic rays striking the earth's atmosphere is roughly $1 \text{ cm}^{-2} \text{ sec}^{-1}$. These particles are mostly protons, but decreasing proportions of heavier atomic nuclei are present ranging from helium (15 per cent of the proton intensity for the same momentum-to-charge ratio) all the way to iron. Although there are many interesting features of the primary composition, one of the most notable is that the abundance of the elements in cosmic rays is very different from the chemical composition of the sun. Not only are cosmic rays relatively rich in heavy nuclei, but there is roughly one million times as much lithium, beryllium, and boron in cosmic rays as in the sun. These light elements presumably arise from collisions of heavy nuclei with interstellar matter. Such considerations determine a value of a few grams per square centimeter for the average amount of matter traversed by cosmic rays before reaching the earth. This, in turn, gives a mean lifetime of cosmic rays in the galaxy (density

$\sim 10^{-26} \text{ g/cm}^3$) on the order of 10^8 years or $\sim 10^6$ years in the spiral arms.

The spectrum of cosmic ray energies between 10^{14} and 10^{19} eV is a power law relation given by $N(>E) = 3 \times 10^{-10} E^{-(1.7 \pm 0.1 \log_{10} E)} (\text{cm}^2 \text{ sec steradian})^{-1}$ where E is the energy in 10^{15} eV. This spectrum should not be extrapolated beyond the indicated upper limit since some experiments indicate a possible change in the slope beyond this point. The maximum particle energies which have been detected are well in excess of 1 joule. These are truly phenomenal energies, equivalent to taking all of the kinetic energy of an apple dropped a distance of several meters and giving it to just one proton of the apple's atoms. Not only are some cosmic rays individually energetic, but because of their high spatial density, cosmic rays represent a large fraction of the total energy associated with astrophysical phenomena. The energy density of cosmic rays, optical photons, interstellar magnetic fields, and the turbulent motions of interstellar matter are each about equal to 1 eV/cm^3 .

The Origin of Cosmic Rays. In the half-century since their discovery by Hess in 1911, the problem of the origin of cosmic rays has remained largely unsolved. The major difficulty has been that of finding a satisfactory mechanism whereby charged particles can be accelerated to the very high energies just discussed. Our sun is wholly inadequate in this respect; the sun fails on all four counts mentioned previously to be considered as the sole source of cosmic rays.

Present-day speculation considers the sources to be violently active celestial objects; exploding galaxies and exploding stars or supernovae. There are several reasons why such objects may indeed be the long-sought sources. For one thing, supernovae are known to be rich in heavy elements. For another, both types of explosive phenomena involve huge amounts of energy. This is inferred from the presence of synchrotron radiation (the emission of electromagnetic waves by electrons moving in magnetic fields) which implies in some instances electron energies as high as 10^{13} eV. These electrons are most likely the decay products of mu-mesons which are again the decay products of charged pi-mesons. The pi-mesons are created in nuclear interactions, thereby suggesting the presence of very energetic nuclei, some of which could escape from the magnetic fields of the supernovae to become cosmic rays. There are possibilities of verifying such a model, for neutral pi-mesons should also be produced in the nuclear interactions. These mesons decay into gamma rays which travel in straight lines unaffected by magnetic fields. Thus, experiments to detect high-energy gamma rays coming from supernovae and radiogalaxies will be of much aid in solving the long-standing problem of cosmic-ray origin.

High-energy photons have already been observed in the primary cosmic rays. Gamma and x-rays produced by interaction of primary particles with interstellar matter and optical photons (inverse Compton scattering) give information on the distribution and composition of

matter in our galaxy. Also, several point sources of x-rays have been discovered, one of which is in the Crab Nebula. The processes whereby these x-rays are created are not yet known.

All high-energy interactions lead ultimately to the production of neutrinos. Although the detection of primary neutrinos will be terribly difficult, neutrino astronomy would be a useful tool for the study of astrophysical phenomena.

Solar Effects. Although the sun has little effect on the high-energy cosmic-ray flux, it strongly influences the low-energy flux during the occurrence of solar flares. At such periods of solar activity protons may be emitted by the sun with kinetic energies of nearly 10^{11} eV. The accompanying increase in the sea-level cosmic-ray intensity can be as much as fifty times larger than the normal value of about $1.0 \text{ cm}^{-2} \text{ min}^{-1}$.

The earth is surrounded by electron ring currents circulating in the earth's magnetic field. These ring currents, which are named Van Allen belts after their discoverer, contain electrons with energies ranging from 10^3 to 10^5 eV. Occasionally, the sun emits an ionized gas which moves outward with a velocity of 10^3 km/sec. These particles do not have enough energy to penetrate the earth's magnetic field, but they do modify the field, thereby releasing electrons trapped in the Van Allen belts. When the electrons strike the atmosphere, the resultant ionization produces auroras. The magnetic fields associated with the ionized gas also prevent low-energy galactic cosmic rays from reaching the earth. The accompanying decreases in the sea-level cosmic-ray intensity are known as Forbush decreases. Since these fluctuations in intensity are related to solar activity, their occurrence is periodic and associated with the 11-year solar sunspot cycle.

The Secondary Radiation. If it were not for the secondary cosmic rays, high-energy primaries would never have been discovered. Consider a particle detector of area 1 cm^2 and aperture 1 steradian. With such a detector one must wait nearly 10^{10} years, the lifetime of the universe, before registering the passage of a 1-joule particle. As we shall now see, the detection frequency is enormously enhanced by the earth's atmosphere.

The layer of air above the earth represents about 13 collision mean free paths for an incident proton. After the inevitable interaction of a primary particle with some atom high in the atmosphere, the nuclear debris so produced undergoes successive interactions with air atoms further down in the atmosphere. In each collision pi-mesons are created which decay as described earlier into mu-mesons and gamma rays. Owing to their large Lorentz factors commensurate with their high energies, the mu-mesons continue on down to sea level before decaying. The gamma rays, on the other hand, produce electron-positron pairs which, in turn, radiate more gamma rays. The huge number of electrons created in this way is called an extensive air shower (EAS). After reaching a maximum at an atmospheric depth dependent on the energy of the primary particle,

the EAS slowly decays by ionization losses. Even so, energetic primaries can give rise to EAS which contain billions of electrons at sea level. The total number of EAS particles reaching sea level is nearly proportional to the primary energy, the relation being roughly 10^9 to 10^{10} primary eV per secondary electron.

As they cascade down through the atmosphere, the electrons are scattered by air atoms so that, upon reaching sea level, the EAS electrons and positrons are distributed over a large area. The density distribution of these secondaries is peaked around the shower axis (the direction of motion of the primary particle) and decreases monotonically with distance in a plane perpendicular to the shower axis. For example, a 1-joule primary can create detectable numbers of electrons per square meter even at distances of 1 km from the shower axis. Thus, a small number of detectors spread out in a plane can (a) detect EAS produced by energetic primaries, (b) give information concerning the primary energy from knowledge of the secondary electron density distribution, and (c) be used to determine the incidence angle of the primary particle. The last measurement involves timing the arrival of the shower front (the nearly plane surface containing the majority of the secondaries and propagating along the shower axis with the velocity of light) at several distances from the shower axis. By spreading a dozen or so detectors over an area of 10^6 square meters, several showers representing primaries of 1 joule or more can be detected per year.

In order to extend the detectable upper limit of primary cosmic-ray energies, other techniques are in use which effectively increase the sensitive area of the individual detectors. These methods involve the detection of the atmospheric fluorescence produced isotropically (hence detectable over a wider area than the electrons themselves) by the secondary electrons as they pass through the atmosphere.

Among the many uses of the secondary radiation has been the discovery of new particles, notably the positron and the various mesons, and the study of their interactions with matter. In addition, some of the interactions provide remarkable clocks for finding the age of many terrestrial features. For example, the collisions of secondary neutrons with atmospheric nitrogen produce carbon-14 which combines with oxygen to form radioactive CO_2 , thus facilitating the familiar technique of radiocarbon dating developed by Libby.

JOHN P. DELVAILLE

References

- Progress in Cosmic Ray Physics*, 1-3, and subsequent volumes entitled *Progress in Elementary Particle and Cosmic Ray Physics*, Amsterdam, North Holland Publishing Co.
- Hayakawa, S., "The Origin of Cosmic Rays," *Lectures on Astrophysics and Weak Interactions II*, Brandeis University (1963).

Chiu, H.-Y., "Neutrino Astrophysics," *Lectures on Astrophysics and Weak Interactions II*, Brandeis University, 165 (1963).

Greisen, K., "Cosmic Ray Showers," *Ann. Rev. Nucl. Sci.*, **10**, 63 (1960).

Ginzburg, V. L. and Syrovatskii, S. I., "The Origin of Cosmic Rays," New York, Pergamon Press, 1964.

Cross-references: ASTROPHYSICS; ELEMENTARY PARTICLES; ELEMENTS, CHEMICAL; RADIATION BELTS; RADIO ASTRONOMY.

COSMOLOGY

Cosmology is the study of the structure and evolution of the universe in the large. The observable universe consists of a hierarchy of stars, galaxies and clusters of galaxies, the latter having typical dimensions of about 1.6 megaparsecs* (Mpc) or 5 million light-years, and containing a few hundred members. It is still in dispute as to whether there are clusters of clusters of galaxies or even higher-order structure. It is generally considered that the space between galaxies and clusters is permeated with a tenuous gas. However, this has not been directly observed and only crude limits can be put on the amount and physical state of the intergalactic medium. The smoothed-out density of the matter which is observable in the form of galaxies is about 5×10^{-30} g/cm³. The total density, including the unobserved intergalactic matter could be two or three orders of magnitude higher than this.

It was discovered about forty years ago that the lines in the spectra of galaxies are redshifted with respect to their terrestrial wavelengths and that the red shift increases linearly with increasing distance. It is generally supposed that this is due to the Doppler effect and that, consequently, the universe is expanding. It is held by a minority that the universe is static and that photons lose energy while traversing the intergalactic medium by physical processes which are not yet understood. The rate at which the apparent velocity of recession increases with increasing distance is known as Hubble's constant, H . Current observations indicate the H is about 100 km/sec/Mpc; this value could be wrong by as much as a factor of two. The most distant known galaxies have redshift $\Delta\lambda/\lambda \approx 0.5$ and lie at distances of roughly 1000 Mpc, or 3 billion light-years.

Theoretical cosmologists investigate the properties of idealized "world models" in which, for example, the observed patchy distribution of matter is replaced by a smoothed-out distribution. In order to construct the mathematical models, simplifying assumptions must be made, and the uniqueness of the universe, coupled with the fact that we have only been able to survey a small part of it, means that these assumptions are only partly based on observation. The most

generally adopted assumption is the "cosmological principle"—that the universe has the same general character as seen from any point within it at a given time. A more powerful assumption, the "perfect cosmological principle," that the universe has the same general character as seen from any point *at any time*, forms the starting point of the steady-state cosmology proposed by Bondi and Gold in 1948. In this theory, matter is supposed to be continually created throughout the universe so as to exactly balance that lost by the observed expansion. Thus, cosmology is not, strictly speaking, a branch of physics because although the prediction of a particular cosmology must be consistent with the laws of physics as we know them locally, the cosmologist does not always start by extrapolating these known laws to the large scale.

Probably the first attempt to apply modern ideas to the universe as a whole was made by Olbers in 1826. Olbers assumed that space is Euclidean, that it is filled with stars whose average number per unit volume and whose average luminosity are constant throughout space and time, and that there are no large-scale systematic motions of the stars relative to one another. He was then able to show that if the universe has infinite extension in space and time, the radiation density at any point should equal that at the surface of a typical star! This result was a paradox to Olbers and the paradox remained unresolved until the discovery of the expansion of the universe one hundred years later.

The first dynamical models of the universe were derived after the publication of Einstein's General Theory of Relativity in 1915. This theory, with its idea that space is only locally Euclidean and that the presence of matter influences the curvature of space, made it possible to devise world models which are isotropic and finite in volume but which have no boundary in the sense that a light ray can travel on and on indefinitely. It is also possible to construct expanding or contracting models in which the density of matter and radiation remains uniform but a function of the time and in which there is no preferred center of expansion. Thus the space between clusters of galaxies is imagined to expand or contract in a manner similar to the behavior of a uniform distribution of points on the surface of an expanding or contracting balloon.

We now describe some of the main world models which have resulted from solutions of the field equations of general relativity. In 1916, Einstein devised a static model of the universe in which the density is uniform and in which space is so curved that the universe has a finite volume. In order to obtain this solution and to overcome certain mathematical difficulties, Einstein had to incorporate an additional term, known as the cosmological constant, into his equations. This term represents a repulsive force between masses which increases with their separation. This term has a negligible effect on terrestrial phenomena and on the motions of the planets, but it can have an important effect on the cosmic

* A parsec is the distance at which the radius of the earth's orbit around the sun subtends an angle of one second of arc. It approximately equals 3×10^{18} cm.

scale and, in Einstein's static universe, the repulsive term just balances the gravitational attraction.

In 1917, de Sitter found a further solution of Einstein's equations for a universe of vanishingly small density. It represented an expanding universe in which test particles of negligible mass would recede from one another with ever increasing velocity. It was just at this time that astronomers were discovering the red shift of the galaxies, and de Sitter's model, although it represented a universe having zero mass, helped to establish the idea of an expanding universe.

In Einstein's static model, the equilibrium between the gravitational attraction and the cosmical repulsion is unstable. Thus a slight contraction of the universe would lead to the gravitational attraction dominating the repulsion and the universe would continue to contract, and vice versa. On this basis, Lemaître and Eddington were able to devise a universe having an infinite past in the unstable Einstein state until some unknown effect started off the expansion which would then go on at an ever increasing rate.

In all the theories, with the exception of the steady-state theory, the age of the universe (i.e., the time which has elapsed since the beginning of the expansion) is of the order of the reciprocal of Hubble's constant. The currently adopted value of $H = 100 \text{ km/sec/Mpc}$ leads to a "Hubble time" of about 10^{10} years. The measurement of the Hubble constant is technically very difficult and depends in essence on measuring the *apparent* luminosity of an object of known *intrinsic* luminosity. Distances to galaxies within a few megaparsecs can be measured by using cepheid variable stars for which there is a relation between the period of light variation and the intrinsic luminosity; this relation can be calibrated using cepheids in our own galaxy. Unfortunately, distances within our own galaxy are difficult to measure accurately; hence, until recently, the adopted value of Hubble's constant was about 500 km/sec/Mpc . This led to a "Hubble time" of less than 2×10^9 years whereas the earth is known to be about 3.5×10^9 years old! In order to avoid a situation in which the solar system is older than the whole universe, cosmological theories have been devised, principally by Dirac, by Jordan, and by Dicke, in which the universal constants (for example, G , the constant of gravitation) are functions of time. The time-scale problem is now less serious, but the astrophysically determined ages of the oldest star clusters in our galaxy come out to be about 16×10^9 years, or almost twice the Hubble time. This discrepancy is generally not considered serious in view of the large uncertainties in the age dating of star clusters and the measurement of Hubble's constant.

However, if the age discrepancy persists it will be strong evidence in favor of the steady-state theory of Bondi and Gold, in which the universe has an infinite past and an infinite future. In this theory, the loss of matter by the expansion of the universe is exactly compensated by the creation of new matter so that the formation of new

galaxies is always going on and only the *average* galaxy need have an age equal to the Hubble time. The steady-state theory has also been vindicated to some extent by recent work on the origin of the elements. It is known that most of the mass of the universe is in the form of hydrogen, the simplest and lightest element, and until about 1950, it was generally considered that conditions necessary for the synthesis of the heavier elements could only have existed in some primeval compact state of the universe before the expansion began. However, it has now been shown that element building can go on inside stars during the normal course of their evolution and that, indeed, the primeval gas out of which our galaxy formed was almost pure hydrogen.

The original form of the steady-state theory was later elaborated by Hoyle who added additional terms to the field equations of general relativity to represent the creation of matter. In 1963, Hoyle and Narlikar advanced a new version of the theory which is claimed to incorporate Mach's principle—the idea that the inertia of a body is due to its interaction with distant parts of the universe. Failure to incorporate Mach's principle has long been considered to be a major drawback to the cosmologies based on the conventional field equations.

On the observational side the major step forward in recent years has been the discovery that many of the radio astronomical sources are extremely distant abnormal galaxies. This has led to the discovery of the most distant known galaxies and to attempts to decide between the various cosmological models by examining the way in which the number of sources increase with decreasing luminosity. The first indications were that the data favors an evolutionary rather than the steady-state cosmology, but this cannot be confirmed until more is known about the intrinsic distribution in luminosity of the radio galaxies.

W. L. W. SARGENT

References

- Bondi, H., "Cosmology," Second edition, London and New York, Cambridge University Press, 1961.
- Hubble, E. P., "The Realm of the Nebulae," New Haven, Conn., Yale University Press, 1936; reprinted by Dover Publications, Inc., New York, 1958.
- Sciama, D. W., "The Unity of the Universe," Garden City, N.Y., Doubleday Anchor Press, 1961.

Cross-references: ASTROMETRY; ASTROPHYSICS; DOPPLER EFFECT; ELEMENTS, CHEMICAL; RELATIVITY; SOLAR PHYSICS; SPACE PHYSICS.

CRITICAL MASS*

The mass of fissionable material required to produce a self-sustaining sequence of fission reactions in a system (a reactor, for example) is the *critical mass* for that system.

* Research sponsored by the USAEC under contract with Union Carbide Corporation.

The chain of reactions will be self-sustaining if, on the average, the neutrons released in each fission event initiate one new fission event. The system is said to be *critical* when that condition exists.

Neutrons released from fissioning nuclei may escape from the system; they may be captured in non-fissioning reactions, or they may produce new fissions. The critical mass depends on the relative probabilities of these processes and on the average number of neutrons released per fission. Evaluation of these probabilities is the concern of criticality calculations which are important in the design of neutron chain reactors.

The escape probability becomes larger for smaller systems, inasmuch as the ratio of surface to volume increases as a system is made smaller. Thus, there is a *critical size* below which the chain reaction in a given system cannot be made self-sustaining. The concept of critical size is often discussed along with critical mass.

Neutrons colliding with non-fissionable nuclei in the system may be absorbed and thus lost to the chain reaction. In fact, not every neutron absorption by a fissionable nucleus results in a fission. Non-fission absorption must be taken into account in calculating the critical mass. For example, a system containing pure U^{235} can be made to have a low critical mass. If the same configuration were loaded with a sufficient quantity of natural uranium (0.0057 per cent U^{233} , 0.72 per cent U^{235} , and 99.27 per cent U^{238}) to contain the same total amount of U^{235} , it would not be critical because at certain energies the U^{238} readily absorbs neutrons without fissioning.

The probability that a neutron striking a fissionable nucleus will cause it to fission depends on the fission cross section (see FISSION) which in turn depends on the energy of the neutron, increasing as the neutron energy gets lower. Thus the addition of a moderator, that is, a material which takes up energy from the neutrons without absorbing them, will lower the critical mass of a system. Water and carbon are good moderators.

The critical mass also depends on the average number of neutrons released per fission. This number changes slightly with neutron energy. For U^{235} it is about 2.45 for thermal neutrons and about 2.65 for 1 MeV neutrons. The numbers are slightly higher for U^{233} .

A complete criticality calculation must take into account the fission cross section as a function of neutron energy and the average neutron yield per fission as a function of neutron energy. Also to be considered are the geometrical distributions in the system of the fissionable nuclei, the absorbing nuclei, and the moderator, and how the neutrons scatter from these. Furthermore, the configuration of reflecting material outside of the fuel volume has a marked influence on the critical mass. A complete calculation would construct the spatial and energy distributions of neutrons in the system through the use of a mathematical procedure that models the history of neutrons from their release to their capture or escape. Neutron diffusion theory or transport theory is usually used

for such calculations. Actual calculations use idealizations and produce approximate results. Often, criticality experiments are required to verify results.

Finally, as examples of critical masses, a sphere 32 cm in diameter containing U^{235} dissolved in water has a critical mass of about 2.1 kg. The same sphere with U^{233} has a critical mass of about 1.1 kg.¹ The Oak Ridge National Laboratory graphite reactor was loaded with 31 tons of natural uranium. This contains about 203 kg of U^{235} .

A thorough discussion of neutron chain reactors is given in reference 2.

C. D. GOODMAN

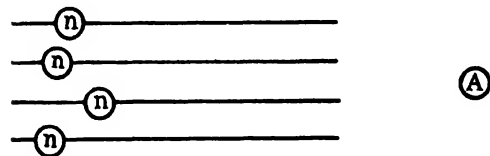
References

1. Callihan, A. D., Morfitt, J. W., and Thomas, J. T., *Proc. Intern. Conf. Peaceful Uses At Energy, Geneva*, 5, 145 (1956).
2. Weinberg, A. M., and Wigner, E. P., "The Physical Theory of Neutron Chain Reactors," Chicago, University of Chicago Press, 1958.

Cross-references: CROSS SECTIONS AND STOPPING POWER; FISSION; NUCLEAR REACTIONS; NUCLEAR REACTORS.

CROSS SECTION AND STOPPING POWER

Cross section, σ , is a conceptual quantity widely used in physics, particularly in nuclear physics, to represent the probability of collision between particles. For example, if a beam of neutrons (n) is incident from the left on a nucleus (A), a certain fraction of the neutrons will be removed from the beam by interaction with A.



By definition, σ is the fraction of neutrons, contained in 1 cm² of beam, that interact with A.

For a thin layer of material of thickness dx containing N nuclei/cm³ the number of nuclei/cm² is $N dx$. If the flux of incident neutrons is ϕ /cm², the fractional decrease in traversing the thin layer will be:

$$-(d\phi/\phi) = N\sigma dx \quad (1)$$

or for a finite thickness x and incident flux ϕ_0 :

$$\phi = \phi_0 e^{-N\sigma x} \quad (2)$$

For a given type of nucleus there are, in general, a number of possible interactions, a, b, c, \dots . The total cross section σ_t is the sum of the cross sections for the individual interactions:

$$\sigma_t = \sigma_a + \sigma_b + \sigma_c \dots \text{etc.} \quad (3)$$

A convenient unit for nuclear cross sections is the *fermi* or *barn*, an area of 10^{-24} cm², which is approximately equal to the cross-sectional area of medium weight nuclei.

The *stopping power*, Sp , of a material for an incident particle is the quantity $-dT/dx$, i.e., the energy loss per unit length of path, generally expressed in ergs per centimeter. This type of attenuation is used primarily in considerations of the passage of heavy charged particles, such as protons, deuterons and alpha particles, through matter. For example, for a particle of any spin having a rest mass $M(\gg m_0)$, the rest mass of an electron), charge ze , and velocity $V(=\beta c)$, the energy loss (as excitation and ionization) per element of path $-dT/dx$ to a homogeneous medium containing N atoms cm⁻³, each of atomic number Z , is given by:

$$-\frac{dT}{dx} = \frac{4\pi e^4 z^2}{m_0 V^2} NZ \left[\ln \frac{2m_0 V^2}{I} - \ln(1 - \beta^2) - \beta^2 \right] \quad (4)$$

where I is the mean atomic excitation potential calculable from the Thomas-Fermi electron distribution function to be $I = kZ \approx 11.5$ electron volts.

The *relative stopping power* S is the inverse ratio of the length of a material to the length of a standard substance having equivalent stopping power (usually referred to aluminum as $S_0 = 1$) at 15°C and 76 cm pressure:

$$S = \frac{(dT/dx)_1}{(dT/dx)_0} = \frac{N_1 B_1}{N_0 B_0} = \frac{\rho_1 B_1 A_0}{\rho_0 B_0 A_1} \quad (5)$$

where ρ is the density, A the atomic weight, and $B = Z \ln 2m_0 V^2 / I$ from Eq. (4).

CLARK GOODMAN

Cross-reference: COLLISIONS OF PARTICLES.

CRYOGENICS

Cryogenics is the production and study of phenomena which occur at very low temperatures, i.e., below about 80°K. The first step in attaining the required temperature generally involves the liquefaction of a gas or gases. Liquids can exist over a range of temperatures limited by the critical point at the higher end and the triple point at the low-temperature end. It is thus possible to compress a gas to the liquid phase at the critical point and to cool it by boiling under reduced pressure to its triple point. A series of gases having their critical and triple points overlapping can thus be used in a cascade process each being used as the refrigerant for the next in the series. Pictet used this method to liquify oxygen, using methyl chloride and ethylene as refrigerants. There are however no liquids which cover the range from 77°K to the critical point of hydrogen or from 14°K to the critical point of helium (5.2°K). Thus liquid hydrogen and helium cannot be produced by the cascade method.

A gas may also be cooled by making it do work in the course of an expansion. When an ideal gas is expanded through an aperture into a constant volume, no work is done, since there are no interactions between the molecules and the molecules themselves occupy no volume. When a nonideal gas is so expanded, however, an amount of internal work ($W = (PV)_{\text{final}} - (PV)_{\text{initial}}$) is done against the intermolecular forces; this work may be positive or negative, resulting in a cooling or heating of the gas. Air is cooled by this Joule-Thomson expansion at room temperature, but hydrogen and helium must be precooled to 90 and 15 K to obtain further cooling upon expansion. Using this method, Kamerlingh Onnes first succeeded in liquefying helium in 1908. Compressed gases may also be made to do external work, for example by expansion against a movable piston. In this case the work is always positive and helium may be cooled and liquefied without any pre-cooling by liquid hydrogen. The marketing of a machine of this type designed by Collins has given a great impetus to low-temperature research.

With liquid helium readily available in the laboratory, research in the temperature range 5 to 0.8 K has become commonplace. By using the isotope of helium He³, it is possible to attain temperatures down to about 0.3 K. Since He³ is relatively rare it is used in small closed systems and is pre-cooled by He⁴ which is at about 1 K. Reduction of the pressure over the liquid He³ can cool it to about 0.3 K since it has a lower boiling point than He⁴. This is about the lowest temperature that can be attained by boiling liquids at reduced pressure, and to reach lower temperatures it is necessary to use magnetic phenomena.

It was pointed out by Debye and Giauque that at 1 K the entropy of paramagnetic salts was still fairly large and, moreover, that it was almost all due to nonalignment of magnetic moments and that the entropy of lattice vibrations was very small. If the electron spins are aligned by application of a magnetic field, the entropy of the salt drops to a low value and the heat of magnetization can be extracted isothermally. The salt can then be thermally isolated and demagnetized adiabatically and its temperature will fall. Temperatures in the order of 0.01°K can readily be reached by this method, and the lower limit for a single demagnetization would seem to be about 10^{-3} K. When demagnetization occurs, the spin system reaches equilibrium temperature in about 10^{-10} seconds. It is found that equilibrium is achieved between the spin temperature and the lattice temperature by spin-orbit coupling in times in the order of a few seconds. Paramagnetic salts have relatively high specific heats at low temperature, and hence the cold salt can be used to cool other bodies; however, making good thermal contact to the cooled salt can be difficult.

Kurti, Simon and Gorter suggested that a further reduction in temperature could be attained if adiabatic demagnetization was performed on nuclear moments rather than the electron spin. The temperature which can be reached in an adiabatic demagnetization is determined by the

point at which the entropy of the system in zero external field decreases sharply with decrease in temperature due to the alignment of magnetic moments, i.e., the point at which the interaction energy μh equals kT (h = internal field, μ = magnetic moment). Since the interaction energies of nuclear moments are much smaller than electron-spin interactions, much lower temperatures should result. The materials used experimentally were metals cooled to 10^{-2} °K by contact with a paramagnetic salt. The thermal isolation of the nuclear spins during demagnetization was achieved naturally by the nuclear spin-conduction electron relaxation time (~ 100 seconds). Thus while the nuclear spins cooled to between 10^{-5} and 10^{-6} °K, the conduction electrons and lattice remained in thermal contact with the cooling salt at 10^{-2} °K. Kurti postulates that cooling of the lattice to 10^{-4} °K might be possible if suitable thermal switches can be devised.

Another method of attaining temperatures below 1°K is to use the fact that the entropy of a superconducting metal is less than that of the metal in its normal state. Quenching of a superconductor by the application of a magnetic field can cause a cooling to about 0.1°K. However, since the specific heat of metals is very small at these temperatures, they are not very suitable for cooling other bodies.

Measurements of low temperatures are usually carried out using secondary thermometers which have been calibrated at certain fixed points previously determined on the absolute scale by a standard instrument. This instrument is generally a constant-volume gas thermometer used at low pressures. When the readings of this instrument are extrapolated to zero pressure, the scale coincides with the thermodynamic scale. In the range 0.8° to 5.2°K, the vapor pressure of He¹ provides the most commonly used secondary scale and it agrees with the absolute scale to within 2 millidegrees over this range. The use of He³ instead of He⁴ increases the coverage to 0.3°K.

Resistance thermometers are useful over a wide range; for example, platinum is used from 273 to 15°K, and carbon covers the range 20 to 2°K and has the advantage of being quite insensitive to magnetic field. The above are examples of secondary standards where the scales are interpolated between fixed points. For temperatures below 0.3°K the susceptibility of a paramagnetic salt can be measured and the temperature calculated by extrapolation from Curie's Law $\chi = C/T$. This method gives true values for the temperatures as long as Curie's Law holds. Beyond this region, it is necessary to perform a thermodynamic cycle to determine the relationship between the magnetic temperature T^* and the absolute temperature T . The method of Kurti and Simon is to demagnetize adiabatically from a known temperature on the absolute scale, using a number of different field intensities, and hence to determine the relationship between T^* and the entropy S over the required range of temperature. Measurement of the amount of heat Q necessary to raise the temperature from T^*_1 to T^*_2 gives the absolute value

of the average temperature T_{12} from the relationship $\Delta Q = T \Delta S$. The heat is generally supplied to the salt by gamma rays, thus ensuring even heating of the sample, a necessary precaution since the thermal conductivity is poor. The problem of nonlinearity does not arise in the case of nuclear spin demagnetization since in this case the susceptibility obeys Curie's Law down to 10^{-7} °K.

One of the most interesting phenomena of cryogenics is that of SUPERCONDUCTIVITY, which was also discovered by Kammerlingh Onnes. When metals are cooled from room temperature, their resistivities decrease and at low temperature, they attain low values which are fairly independent of temperature. Some metals, however, have a critical temperature below which their resistance goes to zero. Such a metal is known as a superconductor. In a single crystal of tin, for example, the transition from normal to superconducting may take place in less than a millidegree. If a current is induced in a loop of a normal metal, it decays with a time constant determined by the inductance and the resistance of the loop. In the case of a superconductor, induced currents have been observed to flow for years without any measurable decrease. Since the resistance of a superconductor is zero, the electric field inside it must be zero; hence, from Maxwell's equation $\text{Curl } \mathbf{E} = -\partial \mathbf{B} / \partial t$, we know that \mathbf{B} cannot change inside a superconductor. In fact it has been shown that \mathbf{B} is zero throughout a superconductor. When a superconductor is placed in a magnetic field, supercurrents flow in such a way as to exclude the magnetic field, hence a superconductor is a perfectly diamagnetic body with $\chi = -1/4\pi$. A superconducting material below its critical temperature can be returned to the normal state by a magnetic field (which may be externally applied or may be due to a current flowing in the superconductor). Another property of superconducting metals is the fact that their thermal conductivity is about 1000 times less than that of the same metal in the normal state. Thus superconductors can be used as thermal switches by quenching them with magnetic fields.

An interesting phenomenon which takes place at low temperature is that of superfluidity in liquid helium. As liquid helium is cooled below 2.18°K, it undergoes a sudden discontinuity of specific heat and a second-order transition to the superfluid state. In this state the viscosity of the helium becomes a function of the method used to measure it. Measured by an oscillating disc method, the viscosity falls from 23×10^{-6} poise just above the transition to 1×10^{-6} poise at 1.3°K. Measured by passage through very fine capillaries the viscosity is very nearly zero in the superfluid state. Hence, it is postulated that there are two coexisting non-interacting fluids, one having the properties of non-"superfluid" helium and the other having virtually zero viscosity and zero entropy. It is interesting to note that the thermal conductivity of superfluid helium is about 2000 times greater than that of copper; this is the result of motion of the entropy-free superfluid

rather than normal thermal conductivity (see SUPERFLUIDITY).

The use of magnetic properties in cooling has already been mentioned. There are several fundamental experiments on the magnetic properties of materials which become possible as a result of the low-temperature environment available. The first of these was discovered by deHaas and Van Alphen in 1930. They found that at low temperatures, the susceptibility of bismuth single crystals rose and fell periodically as the magnetic field was increased. Later work has shown that the periodicity occurs in all metals at low temperatures and is the result of quantization of electron motion perpendicularly to the applied field. This effect has recently been used to determine the Fermi surface of metals.

The method of alignment of nuclear moments has been used to study radioactive decay as a function of nuclear orientation. The aligning field in this case can be either an externally applied magnetic field or an internal crystal field. One of the more striking of these experiments has been the test of the Lee-Yang theory of non-conservation of parity in weak interactions. A third fundamental experiment of interest was the confirmation of London's idea that flux through a superconducting ring is quantized. The ring was in fact a lead tube of 10^{-3} -cm diameter, and it was suspended from a torsion balance. The tube was made superconducting in the presence of longitudinal magnetic fields, and the frozen-in flux was measured and found to be quantized.

In addition to the fundamental aspects of cryogenics, several practical applications of low-temperature phenomena have recently been developed. The most striking of these is the superconducting magnet. To achieve magnetic fields in the order of 70 kilogauss by normal methods calls for the expenditure of about a megawatt of power for as long as the field is maintained. The necessity of removing this power by circulation of cooling fluids adds to the inefficiency. When a superconducting coil is used in the persistent current mode, no energy is needed to maintain fields of this magnitude. These magnets have been made possible by the discovery of alloys such as Nb_3Sn and Nb-Zr which have the capacity to remain superconducting while they are subjected to high magnetic fields and are carrying the current necessary to produce the field. A second practical result has been the development of superconducting circuitry. The basic device in this area has been the cryotron, a switch in which the field resulting from current in a superconductor can quench another superconducting superconductor to its normal state. When made in the form of deposited thin films, cryotrons have switching times less than 10^{-8} seconds. The use of thin-film cryotron circuits represents the first attainment of the all-thin-film integrated circuit.

JOHN L. MILFS

References

Jackson, L. C., "Low Temperature Physics," Fifth edition, London, Methuen, 1962.

Mendelssohn, K., "Cryophysics," New York, Interscience Publishers, 1960.

Cross-references: CONSERVATION LAWS AND SYMMETRY, ENTROPY, HEAT TRANSFER, LIQUEFACTION OF GASES, HEAT, SUPERCONDUCTIVITY, SUPERFLUIDITY.

CRYSTALLIZATION

The forms of natural crystals have been studied by mineralogists for many years and have been classified by symmetry, interfacial angles, perfection of shape, and more detailed criteria. These forms are often related to the molecular structure and growth of the crystals. The equilibrium shape of a crystal is that for which the surface energy is a minimum. Since atomic planes of densest packing usually have the lowest surface energy, these planes predominate in the surface facets of equilibrium crystals, resulting in a correspondence between the atomic structure and shape of the crystal. However, natural and even synthetic crystals rarely have the equilibrium shape, because for crystals larger than about 10μ in dimension, the differences in surface energy between faces are too small to transport enough material over the distances required. Therefore the morphology of crystals is usually determined by the rate of crystal growth, rather than by the equilibrium shape.

A crystal is bounded by those faces whose rate of growth is slowest, since fast-growing faces grow out of existence. Close-packed planes frequently grow most slowly, so even when kinetic factors control the crystal shape there is usually a relation between the faces of a crystal and its molecular structure.

Crystals can be grown from the vapor, liquid, or solid phase. Experiments show that crystals form initially in tiny regions of the parent phase and then propagate into it by accretion of material. Thus nucleation and subsequent growth, rather than uniform transformation throughout the parent phase, is the usual mechanism of crystallization.

The formation of a minute crystalline region in a parent phase below its transformation temperature is driven by the difference between the actual and the transformation temperature; however, this formation is opposed by the necessary creation of surface between the crystal and the parent phase. When the volume of a crystalline embryo fluctuates to a value large enough so that this opposing surface energy is overcome, the resulting crystalline nucleus grows. Crystals almost always nucleate on foreign material, such as container walls or impurity particles, because a foreign surface reduces the surface energy for nucleation.

Since crystallization proceeds by propagation of the nucleus into the parent phase, the surface separating these phases is the site of incorporation of molecules into the crystal. If this surface is perfectly smooth, incorporation of molecules into the crystal is difficult; however, if there is a monomolecular step on the surface, incorporation

will occur preferentially at the step. Such a step contains kinks or jogs, which are the final sites for incorporation. Thus the progress of a molecule from the parent phase into the crystal is: (1) transport through the parent phase to the crystal surface, (2) adsorption onto the crystal surface, (3) movement on the surface to a step (surface diffusion), (4) adsorption onto a step, (5) transport along the step to a kink, and (6) incorporation at the kink. Steps (2), (4) and (6) can involve reorientation and desolvation of the molecules. The rates at which crystals grow can be controlled by any one or several of these steps. The rate of removal of the heat of transformation from the crystallizing interface can also influence the over-all growth rate.

Under certain extreme conditions the surface of a crystal is "rough," so that molecules can be incorporated anywhere on it. However, these conditions are rarely encountered, and normally molecules are incorporated only at steps. If the crystal surface is molecularly perfect, a pillbox of material must be nucleated on it to create a

step, which then grows to another perfect surface. Under these circumstances, continued surface nucleation is required, and growth occurs only below a certain undercooling. However, experimentally, crystals often grow at much smaller undercoolings than this calculated one, so that a continuous source of steps must exist. For this source F. C. Frank^{7, p. 48} postulated a screw dislocation in the crystal that emerges at the crystal surface. This emergent dislocation provides a step pinned at one end, so that as it propagates it winds up into a spiral and is always available for incorporation of molecules. Many spirals have been observed on crystal surfaces; one is shown in Fig. 1. The surface nucleation and screw dislocation mechanisms for crystal growth were definitely confirmed by the elegant experiments of G. W. Sears on the growth of perfect metallic filaments ("whiskers"), metallic platelets, and paratoluidine crystals.⁸

Impurity molecules can modify the morphology and growth rate of crystals by their effects on the relative surface energies and growth rates of

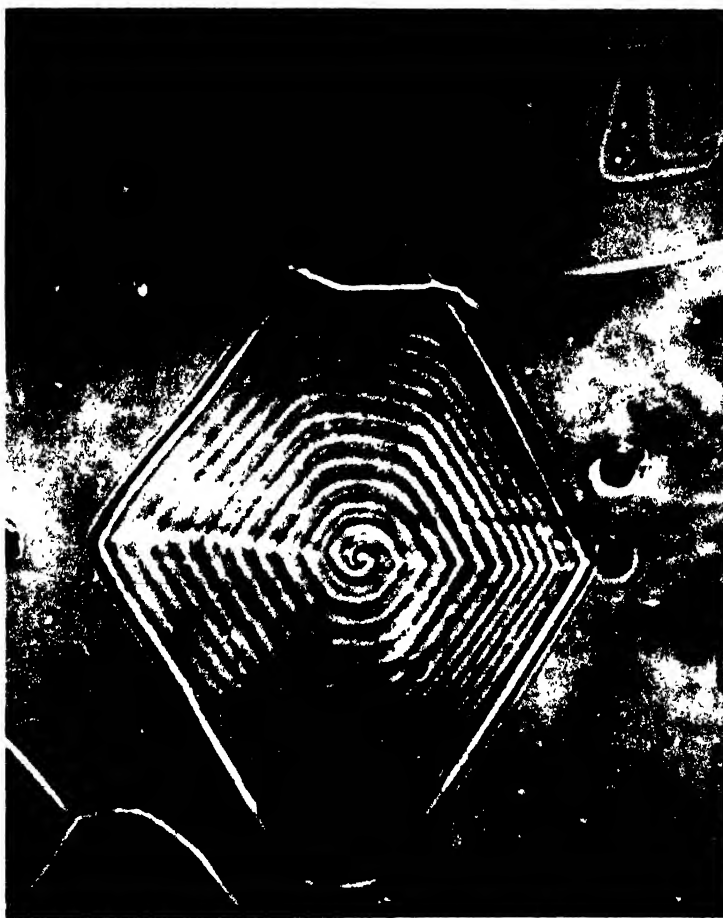


FIG. 1. Growth spiral on a paraffin crystal, observed by C. M. Heck. From Doremus, Roberts and Turnbull "Growth and Perfection of Crystals," by permission of John Wiley & Sons, Inc., New York.⁷

different crystal faces. These molecules can poison growth on certain planes by adsorption at kinks in steps on these planes, slowing the growth of these steps. Impurities can also change the rates of adsorption and surface diffusion of incorporating molecules. The rate of pillbox nucleation can be increased by the lowering of surface tension by impurity adsorption. Therefore impurities can produce different crystal habits and either faster or slower growth rates.

To make crystals for laboratory and industrial use a great variety of techniques have been used. Growth by either condensation or chemical reaction from the vapor phase can give crystals with high purity and special structures and forms. For large-scale industrial use, this method is too costly, although it is valuable for certain special applications. Luminescent crystals of zinc and cadmium sulfides are grown from the vapor for industrial use. Metallic crystals with few impurities and defects, and in the special forms of thin films or whiskers, can readily be grown from the vapor. Other crystals made in this way are silicon, germanium, iodine, selenium, phosphorus, and a variety of organic crystals. The study of the growth of ice crystals from water vapor has special importance in meteorology.

The most common method of growing metallic and semiconducting crystals is by solidification of their melts. Special techniques have been developed to grow single crystals of these materials and many others. In the Czochralski method, a seed crystal is touched to the melt, and the crystal is "pulled" from it by slowly withdrawing the seed. In the Bridgman technique, the melt is slowly moved through a temperature gradient in a furnace, so that crystallization starts at one point in the melt and propagates through it relatively slowly. In the Verneuil method, powder is added to the molten surface of a crystal so that a crucible of other material is not needed. This method is used for materials with high melting temperatures, such as alumina, spinels, rutile, mullite, ferrites, and yttrium-iron garnet. Solid crystals are often purified by zone melting, in which a molten zone is moved through the crystal. Segregation of impurities into the melt purifies the crystal.

Precipitation from liquid solution is a common method of growing crystals. Ionic salts are grown from aqueous solutions both industrially and in the laboratory. Sugar is crystallized from water solution. Other organic crystals, including polymers, are grown from a variety of solvents. Quartz crystals are grown from aqueous solution at elevated temperatures and pressures ("hydrothermal growth"). Various crystals have been grown from more exotic solvents, for example: garnets, titanites, and ferrites from molten salts ("fluxes") tin, iron, and phosphorous from mercury, and diamond from a molten metal under pressure.

Crystallization from the solid phase is also possible. Growth of grain size in a single-phase solid, called recrystallization, is often used to improve the properties of polycrystalline mate-

rials, particularly metals. Crystalline compounds can be made from high-melting materials by pressing together mixtures of their powders and diffusing them together at high temperature ("sintering"). Crystals can be grown from a solid solution. This type of precipitation is frequently used to improve the properties of metals, for example, to harden them.

R. H. DOREMUS

References

1. Holden, A., and Singer, P., "Crystals and Crystal Growing," Garden City, New York, Doubleday and Co., 1960. A simple, nonmathematical discussion of crystal growth and structure.
2. Buckley, H. E., "Crystal Growth," New York, John Wiley & Sons, 1951. A general book with emphasis on growth from solution.
3. Van Hook, A., "Crystallization," New York, Reinhold Publishing Corp., 1961. Another general book emphasizing crystallization from solution, with a discussion of industrial equipment and practice.
4. Gilman, J. J., Ed., "The Art and Science of Growing Crystals," New York, John Wiley & Sons, 1963. The most comprehensive general book, containing articles by different authors on theory, experiment, and practical ways of growing crystals from all phases.
5. Hirth, J. P., and Pound, G. M., "Condensation and Evaporation," New York, The Macmillan Co., 1963. Includes a detailed discussion of theory and laboratory experiments of crystal growth from the vapor, with emphasis on the authors' viewpoints.

Papers from Symposia

6. "Crystal Growth," *Disc. Faraday Soc.*, **5** (1949).
7. Doremus, R. H., Roberts, B. W., and Turnbull, D., Eds., "Growth and Perfection of Crystals," New York, John Wiley & Sons, 1958.

Research paper

8. Sears, G. W., *J. Chem. Phys.*, **24**, 868 (1956) and other references by the same author.

Cross-references: CONDENSATION, CRYSTALLOGRAPHY, VAPOR PRESSURE AND EVAPORATION.

CRYSTALLOGRAPHY

Crystallography is the science of the geometric properties of matter in the ordered solid state. When atoms or molecules condense into a solid phase from a liquid or gaseous phase, the lowest energy state is achieved if they become arranged in as regular a way as possible, usually by forming a small basic unit of structure which is repeated indefinitely in three dimensions throughout the solid to form a *crystal*. The geometric properties of this unit and its manner of regular repetition are highly characteristic of the substance in question, and constitute an exceedingly useful subject of study in connection with any field of science involving the solid state. Occasionally, no extended, regular repetition of structure is present

in the solid phase, but this glassy state has many properties of a liquid and lies outside the realm of crystallography. In other cases, extended order may occur in one direction only (as in fibres) or in two directions (as in some clays), but by far the most common condition of the solid state is full three-dimensional order, and it is with this type of order that crystallography is primarily concerned.

The familiar outward manifestation of the three-dimensional order of the atomic structure of the solid is the polyhedral shape commonly exhibited by crystals. These remarkable shapes were admired for centuries (see, for example, Albrecht Dürer's engraving "Melancholia," 1514), but the underlying principle governing them was first discovered by Steno in 1669. This principle is expressed as the Law of the Constancy of Interfacial Angles, according to which the dihedral angles between the faces of all crystals of a given substance remain unchanged regardless of how the relative sizes and shapes of the faces may vary. René Just Haüy in the late eighteenth century was the first to present a systematic account of the characterization of substances by the measurement of interfacial angles, that is, crystallography, and thus establish it as a science. Haüy was a mineralogist, and through his influence crystallography was subsequently developed and applied by workers mainly in mineralogy, and to a small extent in chemistry. The link between external form and internal structure was dramatically completed by M. von Laue's discovery of x-ray diffraction in 1912, and from then on crystallography was rapidly developed and advanced in physics laboratories, and later more and more in chemistry.

A crystal may or may not exhibit external faces, but if it does, these may be studied in terms of their distribution and development, which constitutes the *morphology* of the crystal, by special techniques of *crystallography*, usually making use of an instrument that reflects beams of light from the crystal faces into a telescope, the *two-circle goniometer*. If the crystal has no faces, its internal geometric properties may be studied by its interaction with radiation, by the methods of *optical crystallography* if refraction of infrared, visible or ultraviolet light is involved, or by *x-ray crystallography* if diffraction of x-rays (also neutrons or electrons) is studied.

The crystal can be defined completely (except for chance irregularities and defects) in terms of the arrangement of the atoms within a finite unit of volume called the *unit cell* (whose size is usually of the order of 10\AA on an edge) and the way this unit is repeated in three dimensions to fill up the volume of the crystal. The shape and dimensions of the unit cell provide parameters characteristic of the substance and constitute the first primary geometrical property of crystals. The most general unit cell (triclinic case) is a parallelepiped which can be defined by six constants, three edge lengths (a , b and c) and three interedge angles (α , β and γ). This unit cell is repeated by *translation*, a shift along each of the cell edges by an integral number of edge lengths. If the unit cell and its contents are

represented by a point in space, the crystal consists of a regular array of such points called a *lattice*, in which each point is related to every other by an integral number of vectorial translations corresponding to the unit cell edges. The lattice should be distinguished from the *crystal structure*, which refers to the arrangement of atoms within the unit cell, although the term "lattice" is sometimes loosely used in reference to the structure.

The atoms within the unit cell may be related to each other by a number of geometric operations called *symmetry*, and this phenomenon constitutes the second chief geometric property of crystals. The unit cell must embrace all of the different types of atoms related by symmetry that are not related by simple translation. On the other hand, the symmetry operations which apply to one unit cell are also operated on by the lattice translations, so that these symmetry operations must also apply to the entire crystal. Thus, the morphology of the crystal and all its other properties must obey this symmetry. The detection and definition of the symmetry also serve to characterize the substance and are equally as important in crystallography as the measurement of the lattice parameters.

The way in which symmetry operations can interact consistently with each other is strictly limited by the geometry of coincidence and can be rigorously analyzed by the mathematical methods of group theory, both as to what symmetry operations are possible and how they may be combined. The problem is usually approached by constraining all symmetry operations to pass through a single point in space, but special restrictions are introduced by the requirement that this point must be consistent with any point in the crystal lattice, that is, the symmetry groups must be consistent with the translational operations of the lattice. An important symmetry operation is the axis of rotation by which any motif is reproduced by a rotation around an axis of $360/n$ degrees, where n is the order of the axis. n successive operations then superimpose the object on itself. In crystals, because of the requirements of the lattice, n can only have the values 1, 2, 3, 4 and 6. A 5-fold axis is not possible, for example, for the same reason that it is not possible to fit regular pentagons into a regular two-dimensional pattern which will fill all space. Further, there are only 11 ways to combine the 5 axes together at a point; these are called the 11 axial point groups. These form a convenient basis for classifying all crystals into 6 *crystal systems*, which, while not strictly rational in their definition, provide a fundamental link between the symmetry of the crystal and its dimensional properties. Reference axes are generally chosen parallel to lattice translation directions, of course, but further, they are customarily taken parallel to prominent symmetry axes. Four of the axial point groups, for example, have a single 3-fold axis or 6-fold axis, with or without a number of 2-fold axes at right angles to them. In all these, the reference c axis is customarily set parallel to

TABLE 1. THE SIX CRYSTAL SYSTEMS

System	Independent Lattice Parameters	Axial Relationships	Number of Symmetry Groups			
			Axial Point Groups	Point Groups	Bravais Lattices	Space Groups
Triclinic	$a, b, c, \alpha, \beta, \gamma$	none	1	2	1	2
Monoclinic	a, b, c, β	$\alpha = \gamma = 90^\circ$	1	3	2	13
Orthorhombic	a, b, c	$\alpha = \beta = \gamma = 90^\circ$	1	3	4	59
Tetragonal	a, c	$b = a$				
		$\alpha = \beta = \gamma = 90^\circ$	2	7	2	68
Hexagonal	a, c	$b = a$				
		$\alpha = \beta = 90^\circ$	4	12	2	52
		$\gamma = 120^\circ$				
Cubic (isometric)	a	$b = c = a$				
		$\alpha = \beta = \gamma = 90^\circ$	2	5	3	36

the unique 3- or 6-fold axis, and the other two axes a_1 and a_2 , which are equivalent by symmetry, are taken normal to the c axis along lattice directions 120° apart, coincident with the 2-fold axes if present. These groups are all included in the hexagonal system.

The axial symmetry operations are operations of the first kind, that is, they reproduce left-hand motifs as left-hand motifs. Other symmetry elements of the second kind, that is, which reproduce left-hand motifs as right-hand motifs, are the center of symmetry and the mirror plane of symmetry. When these operations are added to the axial groups, 32 *point groups* are produced.

Referring to the lattice, the introduction of symmetry gives rise to a number of lattice groups in which various rational relationships exist between the lattice parameters. One important result of this interaction is the appearance of *centered lattices*, in which the lattice unit cell chosen according to the rules used to set up the 6 crystal systems contains additional lattice points on body or face diagonals. There are 14 such "Bravais lattices." The symmetry groups so far mentioned can in favorable circumstances all be detected from the external morphology of the crystal.

When the combinations of lattice translations and symmetry operations, that is, the symmetry properties of the crystal structure, are analyzed, new symmetry operations are evolved (screw axes and glide planes) and each of the point groups contains many such combinations, adding up to a total of 230 *space groups*. These are detected by diffraction methods. Table 1 summarizes the relations between the crystal systems and the symmetry groups.

HOWARD T. EVANS, JR.

References

- deJong, W. F., "General Crystallography," San Francisco, Cal., W. H. Freeman and Co., 1959.
 Phillips, F. C., "An Introduction to Crystallography," New York, Longmans, Green and Co., 1946.

Terpstra, P., and Codd, L. W., "Crystallometry," New York, Academic Press, 1961.

Buerger, M. J., "Elementary Crystallography," New York, John Wiley and Sons, Inc., 1956.

Cross-references: CRYSTALLIZATION, DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS

CYBERNETICS

Cybernetics can be described as "the science of control in machines and animals," but the connotation of the word (or its equivalent in other languages) shows a difference of emphasis between *control* in the United States and the *handling of information* in continental Europe. For the science of control involves three streams of work: closed-loop feedback systems (sometimes called goal-seeking systems); the manipulation of the information which guides these systems; and processes of filtering to exclude casual disturbances from the information channel as far as possible. The second of these, obviously, and the third, less obviously, can lead on to a consideration of communication, control and thought processes in animals.

One of the earliest applications of closed-loop feedback was to the power steering of ships, and a successor to this application is the power-aided steering of automobiles (see *FEEDBACK*). As soon as the steering wheel is turned, producing a temporary discrepancy between the setting of the steering wheel and the angle of the road wheels, power is applied to the steering linkage in the appropriate direction and maintained until the two agree once more. The qualification "as soon as" is important, because if there were a noticeable delay in response, conditions might have changed so that the position originally demanded would be incorrect by the time it was reached. It can be shown mathematically that the occurrence of delay within the closed loop may cause the system to oscillate about the desired position without coming to rest ("hunting"); moreover, there is always a tendency for the magnitude of the delay round the loop to increase with the amount of

amplification in the loop; and hence a high-gain loop (one designed to provide a large driving power in response to a small input signal) is particularly susceptible to this type of instability. The general principles involved for linear systems have been outlined by Bode,¹ and one of the tasks of the "servo" designer is to ensure by analysis (for which there are several mathematical techniques) that the system is free from this type of instability. In practice, his task is complicated by the effects of nonlinearity and of interference or "noise".

Closed-loop control systems are often designed to respond to signals which may be regarded as a kind of "information." For example, in the push-button controlled elevator the passenger informs the machine of the floor to which he wishes to travel by pushing a button, and the elevator continues in motion until the desired floor is reached; or in the programmed control of machine tools, information may be fed in via punched cards, punched paper tape or magnetic tape. "Information" is then an essential part of cybernetics, and the idea of information as a quantity which can be measured in physical terms dates effectively from Shannon's work on the information-capacity of communication channels.² The essence of Shannon's work, the germ of which lay unperceived in the earlier work of the telegraphists Kelvin, Carson and Nyquist, is that information can be defined as a quantity which is invariant to transformations of the wave forms by which it is conveyed. Such transformations are known as "coding" of information. Shannon showed further that there is a maximum rate (to be achieved by "ideal coding") at which information can be communicated without error through a physical channel of specified bandwidth and signal-to-noise ratio. To be measurable, the information must be finite. If there is a finite group of n different signals of which the i th occurs with probability (or relative frequency) p_i , the quantity

$$I = - \sum_{i=1}^n p_i \log p_i$$

is the information conveyed by the group of signals. There are physical as well as mathematical arguments for stating that information is the negative of ENTROPY, the latter quantity being as defined in STATISTICAL MECHANICS. The term cybernetics is sometimes taken to include the handling of information in Shannon's sense of that word, with emphasis on the use of information in control systems rather than on the communication of information over long distances.

The third line of work in cybernetics is the study of methods of minimizing interference, of separating signals from noise, and the classic work in this field is that of Norbert Wiener. Following his work on generalized Fourier series,³ Wiener showed⁴ the relationship between the autocorrelation function of a random waveform and its power spectrum, the Wiener-Khinchine theorem:

$$W(f) = 2 \int_0^\infty \psi(\tau) \cos 2\pi f\tau \, d\tau$$

where $W(f)$ is the power spectrum (spectrum of squared amplitudes) and $\psi(\tau)$ is the autocorrelation function for lag τ . He also showed how related mathematical techniques could be used to design a filter which for given statistical characteristics of signal and of noise will give an output having least mean square error (the Wiener-Hopf equation). Similar mathematical techniques are widely used to determine the internal characteristics of automatically controlled systems (e.g., chemical process plants) without interrupting their operation: either a small random disturbance is superimposed on the control signals, or naturally occurring disturbances are utilized, and comparison of the signals at the input and output of the system allows its internal characteristics to be evaluated.

From the evaluation of the characteristics of a system one proceeds to their improvement and there is a whole class of *adaptive control systems* whose characteristics are automatically modified in accordance with current conditions. One application is to the adjustment of the filtering characteristics of a control system to suit the types of signal and noise being applied to it at any time. Another application is to the optimization of performance of plant such as chemical process control and distillation plant. In this type of plant the function of automatic control was originally to maintain specified conditions of temperature, pressure, flow, etc.; but in optimizing control, the automatic equipment finds and maintains optimum conditions which it can vary in response to changes in external circumstances such as change in the raw material at the input, and it can be used to find experimentally optimum conditions which are not known in advance.

Adaptive systems are more sophisticated than automatic controls having fixed characteristics, and in addition to their superficial resemblance at least to the instinctive behavior of animals, they have other relationships with biological systems, the two most remarkable characteristics of which are the ability to learn and the ability to transfer functions from one physical channel to another in case of damage. It must first be said that the idea of coding of information has made a decisive contribution to knowledge of the mechanism of the nervous systems of animals. Much is now known about the forms of pulse code by which information is sent along nerve channels, and the neuron of biology is coming to be regarded as the prototype element for performing the operations of Boolean algebra, so that the designer of digital computers uses the neuron symbol to denote schematically an elementary circuit which will be constructed from electronic components. The adaptive control system shows some of the characteristics of learning, and electronic models of the type known as "conditional probability machines"⁵ can be made to exhibit behavior resembling that of conditioned reflexes in animals. Attempts have been made to press the analogy further in two directions. Ross Ashby devised a model which he called the "homeostat" which made a random search to find a stable position

which it could then maintain constant, and the constancy bears some analogy to the homeostasis (constancy) of temperature, chemical constitution of blood stream, etc. in the higher animals. Then various workers, such as W. G. Walter⁶, constructed mechanical toy animals which could apparently exhibit elementary forms of animal behavior such as the search for "food," and of which the more sophisticated versions could learn to thread a simple maze.⁷ The relevance of the behavior of these devices to the mechanisms of animal behavior is debatable. Finally, familiarity with the concept of coding of information makes it easier to understand "the genetic code," i.e., the way in which genetic information can be stored in the pattern of chemical groups along the length of a complex molecule, and in some ways, this is the most spectacular contribution of cybernetics to human knowledge.

D. A. BELL

References

1. Bode, H. W., "Network Analysis and Feedback Amplifier Design," Princeton, N.J., D. Van Nostrand Co., 1945.
2. Shannon, C. E., "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, 27, 379 and 623 (1948).
3. Wiener, N., "The Fourier Integral and Certain of its Applications," Cambridge, Cambridge University Press, 1933.
4. Wiener, N., "The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications," New York, John Wiley & Sons, 1949.
5. Uttley, A. M., "Conditional Probability Machines and Conditioned Reflexes," in "Automata Studies," Princeton, N.J., Princeton University Press, 1956.
6. Walter, W. G., "An Electromechanical 'Animal'," *Discovery*, 11, 90 (March, 1950).
7. Deutsch, J. A., "The Insightful Learning Machine," *Discovery*, 16, 515 (December, 1955).

Bibliography of Introductory Books

- Bell, D. A., "Information Theory and its Engineering Applications," London, Pitman, 1962.
- Porter, A., "An Introduction to Servo-Mechanisms," London, Methuen, 1953.
- Bellman, R., "Adaptive Control Processes: a Guided Tour," Princeton, N.J., Princeton University Press, 1961.
- Laning, J. H., Jr., and Battin, R. H., "Random Processes in Automatic Control," New York, McGraw-Hill Book Co., 1956.
- Anfinsen, Christian B., "The Molecular Basis of Evolution," London, Chapman & Hall, 1959.

Cross-references: BIONICS, FEEDBACK.

CYCLOTRON*

The cyclotron is an accelerator of ions widely used to study the nucleus, to produce radioactive substances, and to study the interactions of ion-

izing radiation with living systems and with inert matter. It is equally important as the first of a class—*Magnetic Resonance Accelerators*—which includes the various kinds of synchrotron (see SYNCHROTRON) as well as synchrocyclotrons and sector focused cyclotrons. The essential feature of this type of accelerator is that acceleration of charged particles to high energies is achieved by a successive application of small accelerations in synchronization with the rotational period of the particles in a magnetic field. The condition for synchronization is simple and can be derived as follows: A charged particle moving perpendicularly to the lines of force in a magnetic field will describe a circle which is defined by the equilibrium between the Lorentz force $F_L = eBv$ and the centrifugal force $F_c = mv^2/r$. Equating these gives the rotational frequency of the particles which is set equal to the frequency of the accelerating field. This is the *Cyclotron Resonance Condition*:

$$f_a = f_0 = \frac{eB}{2\pi m} \quad (1)$$

where

- f_a frequency of accelerating field
- f_0 rotational frequency
- e - charge of ion
- m - mass of ion
- B - magnetic field strength.

The important fact is that the rotational frequency is independent of the energy of the particle and depends only on quantities which are (approximately) constant. In 1929 the possibility of using this relationship as the basis for an accelerator occurred to Ernest O. Lawrence, who like many other physicists at the time had been inspired by Rutherford's success in disintegrating atoms with alpha particles from natural sources to seek a means of producing a controlled beam of high energy particles. The practicability of the idea was demonstrated, and most of the essential features of the Cyclotron were developed by Lawrence, M. Stanley Livingston, N. E. Edlefsen, and others during the next few years.

Figure 1 is a schematic diagram showing the principle components of a cyclotron. The dees are two hollow semicircular electrodes in a vacuum tank located between the poles of an electromagnet which provides an approximately uniform magnetic field over the entire region. The dees are part of an electrical resonant circuit which may be excited by an oscillator whose frequency is adjusted to the rotational frequency given by Eq. (1). Ions are produced by an electric discharge in a source located at the center. They are drawn from the source and accelerated into a dee while it is negative, they follow a semicircular path in the (electrostatic) field free interior of the dee and again arrive at the gap between the dees where, by that time, the voltages are reversed in sign and they are accelerated again. The ions describe semicircles of increasing radius as their

*Effort supported by U.S. Atomic Energy Commission.

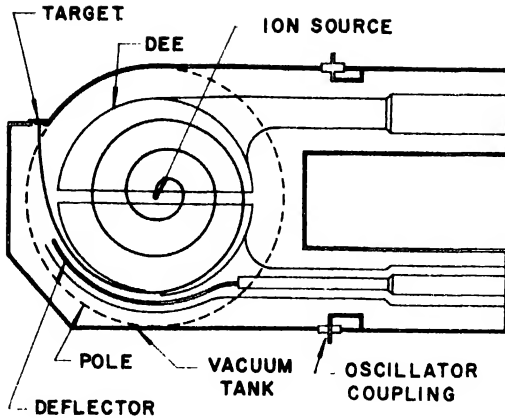


FIG. 1

velocity and energy increase as a result of repeated accelerations. When they reach the maximum radius of the dee, they enter a channel between a septum in one of the dees and the deflector. The deflector is charged negatively and draws the particles out where they may strike a target in the target chamber or they may travel some distance as a beam outside the cyclotron before they are used.

The kinetic energy of the accelerated particles is given by:

(2)

where T is the kinetic energy and R is the radius of ion path at point of extraction. For protons, Eq. (1) reduces to $f = 1.52B$ Mc/sec and Eq. (2) to $T = 3.12 \times 10^{-4} B^2 R^2$ MeV with B in kilogauss and R in inches. The usual values of B are from 15 to 22 kilogauss.

In addition to the resonance condition, a successful cyclotron requires that the orbits be stable, i.e., they must remain in the median plane and at the appropriate radius. The first is achieved by introducing in the magnetic field a small negative gradient with respect to radius. The field lines are then bowed as shown in Fig. 2, and the Lorentz force on a particle off the median plane has a

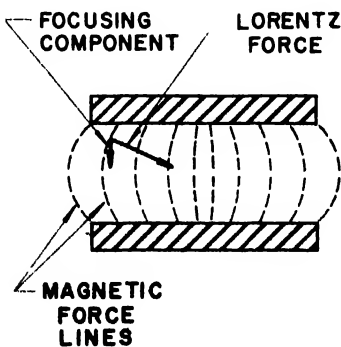


FIG. 2

vertical focusing component. Radial stability results from the fact that the orbit of the particle is an equilibrium orbit with the inward Lorentz force predominating at radii larger than the equilibrium orbit and the centrifugal force predominating at smaller radii. Ions which are displaced either vertically or radially then execute oscillations about the equilibrium orbit. If the magnetic field is described by the index,

$$n = -\frac{r}{B} \frac{\partial B}{\partial r} \quad (3)$$

where r is the radius, then it can be shown that for $0 < n < 1$ stable oscillations occur with frequencies:

$$f_r = f_0 \sqrt{n} \quad (4)$$

$$f_z = f_0 \sqrt{1-n}$$

where f_r is the frequency of vertical oscillations and f_z is the frequency of radial oscillations.

The negative gradient in the magnetic field results, however, in the situation where the rotational frequency, Eq. (1), is not exactly the same at all radii. In addition, it must be noted that the mass in Eq. (1) is the relativistic mass, $m = m_0 / \sqrt{1 - v^2/c^2}$ and increases with energy. The result of these two discrepancies is that the rotational frequency of the ion decreases as it is accelerated and there is an accumulated phase lag between the ion and the accelerating field which when it approaches π radians, results in no further acceleration. The energy limit of the conventional cyclotron can be shown to be proportional to the square root of the accelerating potential and to be about 30 MeV for protons with a dee-to-dee potential of 200 kV. It has not been possible to reach the theoretical maximum energy in practice, and for reasons made clear in the next sections, the incentive to do so has disappeared. The maximum energy which has been attained with protons is 22 MeV and that required about 500 kV on the dees. Currents in cyclotrons are usually of the order of 100 μ A but up to 1 mA has been attained. The most commonly used ions are protons, deuterons, and alpha particles, although heavier ions such as carbon, nitrogen and oxygen ionized to +3 or +4 have also been accelerated.

A possibility of achieving higher energies with cyclotrons was opened up in 1945 when V. Veksler and E. M. McMillan independently pointed out the phase stable characteristic of the cyclotron resonance condition [Eq. (1)] which is apparent if the equation is rewritten in terms of particle energy:

$$f_0 = \frac{Be c^2}{2\pi(E_0 + T)} \quad (5)$$

where E_0 is the rest energy of the particle.

Consider a particle rotating in a cyclotron at the resonant frequency and crossing the accelerating gap at a phase such that it gains no energy and that a later arrival causes it to lose energy. If this

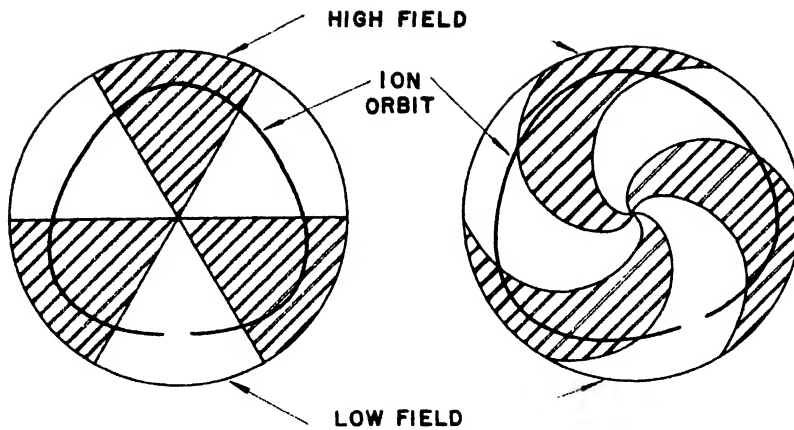


FIG. 3

particle is perturbed by an excess of energy, f_0 decreases and the particle loses energy. If the particle is perturbed in phase so that it arrives at the accelerating gap too early, it gains energy, f_0 decreases, and the phase slips back. Perturbations in energy or phase thus result in oscillations about the equilibrium phase. Under these conditions, if the accelerating frequency of a cyclotron is slowly decreased, the ions will execute stable oscillations about that phase which will give sufficient energy gain so that the radius and energy are matched as the orbits expand. This is the *Principle of Phase Stability* as applied to the Synchrocyclotron. It completely removes the energy limitation of the cyclotron previously discussed. This principle was immediately exploited and synchrocyclotrons (also sometimes called Frequency Modulated or FM Cyclotrons, and in the U.S.S.R., Phasotrons) have been built

which give protons up to 700 MeV. The only limit is the economic one due to the large size of the magnet.

The important structural difference between a synchrocyclotron and a conventional cyclotron is in the provision for a variable frequency. This is accomplished by placing a variable capacitor in the resonant dee circuit. Rotary blade capacitors have been in common use for this purpose, but in more recent designs, vibrating blade capacitors have been preferred. The required frequency swings are about two to one, and the usual modulation frequencies are about 60 to 100 cps. The ions are accelerated in pulses as the accelerating frequency sweeps through its modulation cycle in contrast to the continuous acceleration in a conventional cyclotron. The result is that average currents in synchrocyclotrons are about 1 per cent of cyclotron currents, thus

TABLE I.

	Cyclotron 60-in. Cyclotron, University of Washington, Seattle	Synchrocyclotron 184-in. Cyclotron, Lawrence Radiation Laboratory, Berkeley	Sector Focused Cyclotron 88-in. Cyclotron, Lawrence Radiation Laboratory, Berkeley
Magnet	60-in. pole diameter, 19-kilogauss maximum field, 218 tons	189-in. pole diameter, 23.4-kilogauss maximum field, 4300 tons	88-in. pole diameter, 3 sector, 17-kilogauss maximum average field, 300 tons
RF	Fixed frequency, 11.6 Mc/sec, two dees, 250 kV	Variable frequency, 18-36 Mc/sec for protons, modulation frequency 64 cps, single dee, 9 kV	Fixed frequency, adjustable 5.5-16.5 Mc/sec for various particles, single dee, 70 kV
Beam	Protons 11 MeV, deuterons 22 MeV, alpha particles 44 MeV, 1-mA maximum current	Protons 730 MeV, deuterons 460 MeV, alpha particles 910 MeV, 1- μ A maximum average current	Protons 50 MeV, deuterons 65 MeV, alpha particles 130 MeV, 1 mA maximum current

removal of the energy limit has been accomplished at the expense of a current limitation.

Another method of circumventing the energy limit of the cyclotron was proposed by Thomas in 1938, seven years before the principle of phase stability was enunciated. In the Thomas proposal, the average magnetic field increases with radius so that the resonance condition may be exactly matched by a constant accelerating frequency as the ion gains energy. The axial focusing force is supplied by an azimuthal variation of the magnetic field which may be obtained by using sector magnet poles, Fig. 3. The ion orbits are then no longer circular, and the radial component of velocity interacting with the azimuthal component of magnetic field produces an axial focusing force. This is the "edge focusing" which occurs when an ion crosses a fringe field obliquely, and it has long been used in mass spectrometers and other devices.

This idea was well in advance of the theory and practice of the cyclotron art at the time and was not immediately exploited. Development beginning in 1949 and extending to recent years has resulted in a whole subclass of cyclotrons characterized by a fixed rotational frequency and focusing forces derived from spatial variation in the magnetic field. For example, if the sectors are spiral shaped as in Fig. 3, additional focusing forces of alternating gradient type are developed. These cyclotrons are variously called Sector Focused, Isochronous, and AVF (azimuthally varying field) cyclotrons. They have energies well beyond the energy limit of the conventional cyclotron and, at the same time, are capable of high average currents because they operate at a constant frequency. Provision of auxiliary magnet coils on the pole tips, to trim the field shape over a range of values of average field, and adjustable frequency oscillators, to provide for different ions and a variation in maximum energy, have made the modern cyclotron of this type very flexible.

Table I gives a comparison of the salient features of examples of each of the three main types of cyclotron.

EDWARD J. LOFGREN

References

- Livingood, John Jacob, "Principles of Cyclic Particle Accelerators," Princeton, N.J., D. Van Nostrand, 1961.
- Livingston, M. Stanley, and Blewett, John P., "Particle Accelerators," New York, McGraw-Hill Book Co., 1962.
- McMillan, E. M., in Segrè, Emilio, Ed., "Experimental Nuclear Physics," Vol. 3, pp. 639-785, New York, John Wiley & Sons, Vol. III, 1953-59.

Cross-references: ACCELERATORS, LINEAR; ACCELERATORS, PARTICLE; BETATRON; SYNCHROTRON; ACCELERATOR, VAN DE GRAAFF.

CYCLOTRON RESONANCE (DIAMAGNETIC RESONANCE)*

The term cyclotron resonance is used to designate the resonant coupling of electromagnetic power into a system of charged particles undergoing periodic orbital motion in a uniform static magnetic field. The frequency of the electric field at resonance is simply related to the orbital frequency of the electron in the magnetic field. The effect has been observed and studied extensively in gases and in solids.

One important application of the cyclotron resonance principle is made in the acceleration of charged particles, as in a cyclotron. In a uniform magnetic field, H , a charged particle of mass, m_e , undergoes orbital motion with an angular velocity

$$\omega_e = \frac{eH}{mc}, \quad (1)$$

in which e is the charge and c the velocity of light. Energy from the electromagnetic fields, i.e., from the alternating electric and magnetic fields, is transferred into kinetic energy of the particle, and the radius of the particle orbit is increased with no change in angular velocity. Particle acceleration takes place *in vacuo* in order to prevent energy transfer to the gas by means of collisions.

In solids, cyclotron resonance has been successfully applied to studies of electronic energy band structure. The perfectly periodic array of atoms in an ideal solid scatters electrons coherently. An electron experiencing such coherent scattering can be described by the same equations of motion as the free electron, except that the free electron mass is replaced by an effective mass, m^* . Incoherent scattering from crystalline imperfections causes electronic collisions which limit the number of completed electron orbits, thus giving rise to a frequency bandwidth for the cyclotron resonance absorption. The observation of cyclotron resonance requires that the charged particle execute about one complete cyclotron orbit without collisions, or $\omega_e \tau \gtrsim 1$, in which the collision time, τ , is the mean time between incoherent scatterings. A long collision time is achieved by using samples of the highest possible purity and lattice perfection and by cooling to very low temperature (usually liquid He temperature, 4°K) to eliminate the thermal motion of the atoms. The condition for cyclotron resonance can also be satisfied by increasing ω_e through the use of high magnetic fields, e.g., 100-kilogauss static fields are currently available which for free electrons results in $\omega_e \approx 1.5 \times 10^{12}$ rad/sec or an electromagnetic wave length of about 1 mm.

Electrons moving in the periodic lattice of a solid occupy energy levels which are specified by the wave vector quantum number, k , or by the crystal momentum, $\hbar k$. Since the number of electrons is very large, the wave vectors assume an

* Support of U.S. Air Force is acknowledged.

almost continuous range of values. A knowledge of the functional form of the dependence of the energy on wave vector is necessary for a complete description of the behavior of electrons in solids. The simplest form of the relation between energy and wave vector valid for energy bands in cubic crystals is

$$E(\mathbf{k}) = \frac{\hbar^2 k^2}{2m^*}. \quad (2)$$

In this case, the constant energy surfaces in wave vector space are spheres, and the cyclotron mass of Eq. (1) is just the effective mass, m^* . For energy extrema located at general points in wave vector space, Eq. (2) becomes

$$E(\mathbf{k}) = \frac{\hbar^2}{2} \left(\frac{k_x^2}{m_x} + \frac{k_y^2}{m_y} + \frac{k_z^2}{m_z} \right) \quad (3)$$

in which the extremal point is taken as the origin, and m_x , m_y , and m_z are three components of an effective mass tensor. This generalization is also necessary in describing the energy bands for crystals with symmetry lower than cubic.

An expression for the cyclotron effective mass which is valid for an electron orbiting on a constant energy surface of energy E for an arbitrary $E(\mathbf{k})$ is

$$m_c(E, k_H) = 2\pi\hbar^2 \left(\frac{\partial A}{\partial E} \right)_{k_H} \quad (4)$$

in which k_H is the wave vector component parallel to the magnetic field, A is the area of the electron orbit in wave vector space, and $(\partial A / \partial E)_{k_H}$ is the derivative of this area with respect to energy evaluated at constant k_H . For spherical or ellipsoidal constant energy surfaces, the cyclotron mass is independent of both energy and k_H , and for these two simple cases, m_c is given, respectively, by $m_c = m^*$, and

$$\left(\frac{1}{m_c} \right)^2 = \frac{\alpha^2}{m_y m_z} + \frac{\beta^2}{m_x m_z} + \frac{\gamma^2}{m_x m_y} \quad (5)$$

in which α , β , γ are the direction cosines of the magnetic field with respect to the axes of the ellipsoidal constant energy surface.

For solids which have relatively low carrier density (e.g., insulators, semiconductors, and semimetals) the electronic states which are important in the transport properties are located near energy band extrema. For nondegenerate extremal points in wave vector space, $E(\mathbf{k})$ can be expanded in a Taylor's expansion. The leading term of such an expansion would be given by Eq. (3). For degenerate points (positions where two or more levels have the same energy), a simple generalization of a Taylor's expansion must be used. In solids with relatively high carrier density (e.g., metals), the transport properties are determined by electronic states which are far from the energy extrema and the $E(\mathbf{k})$ relation is not adequately described by a Taylor's expansion.

Cyclotron resonance experiments have been particularly successful in the quantitative determination of the band parameters of the semiconductors silicon and germanium. The successful application of this technique in these semiconductors is attributed to the high quality of the available material, and to the complete classification of the possible forms of the theoretical band structure model. Since in these materials the intrinsic carrier concentration is extremely small at low temperatures, electrons are optically excited out of filled valence levels in the crystal in order to produce sufficient carriers to obtain a measurable signal. Resonances are observed both for the excited electrons and for the holes left behind in the empty levels in the valence band.

In metals, the high carrier density requires modification of the conventional cyclotron resonance experiment. Two important consequences of this high carrier density are the nonuniform penetration of the electromagnetic field in the skin depth and the inapplicability of the simple effective mass theory to describe the electronic states. To overcome the problem of the small electromagnetic penetration depth, the geometrical arrangement suggested by Azbel and Kaner is used. The static magnetic field is applied in the plane of a flat sample, so that the electrons near the surface can be accelerated by the electromagnetic fields, and the orbits described by a cyclotron radius which is large compared with the skin depth. In this way, whenever the applied frequency is a multiple of the cyclotron frequency, a resonant condition is satisfied. This type of cyclotron resonance experiment yields an effective mass at the Fermi energy given by Eq. (4), which is, in general, dependent on the wave vector component parallel to the magnetic field. The interpretation of these experiments is not simple but when coupled with experiments which measure the shape of the Fermi surface, such as DE HAAS-VAN ALPHEN EXPERIMENTS, a fairly complete determination of the electronic band structure is possible. These techniques have been successfully applied in the study of copper.

Cyclotron resonance in ionic crystals allows the measurement of polaron effects. The POLARON denotes the charge carrier together with its local lattice distortion. Cyclotron resonance observed in AgBr has been interpreted as a polaron orbiting in the applied magnetic field.

G. DRESSLHAUS

References

- Kittel, C., "Introduction to Solid State Physics," sec. ed. p. 371, New York, John Wiley & Sons, Inc., 1956.
- Lax, B., and Mavroides, J. G., "Solid State Physics," Vol. XI, p. 261, New York, Academic Press Inc., 1960.

Cross-references: CYCLOTRON, DE HAAS-VAN ALPHEN EFFECT, ENERGY LEVELS.

D

DE HAAS-VAN ALPHEN EFFECT

In 1930, W. J. de Haas and P. M. van Alphen observed an anomalous behavior in the magnetic susceptibility of a pure bismuth single crystal at very low temperatures. Subsequent studies showed that the susceptibility oscillated with changing magnetic field and was in fact periodic in the reciprocal field. This effect was thought for some years to be peculiar to bismuth; however, in 1947 it was found in zinc and, shortly afterwards, in a number of other metals. The effect has now been observed in most metals (including ferromagnetic elements), as well as intermetallic compounds and ordered alloys. Related magneto-oscillatory behavior is observed in the electrical resistance, Hall effect, specific heat, ultrasonic attenuation and velocity, and the thermoelectric power. While originally thought to be somewhat of a scientific curiosity, the de Haas-van Alphen effect has become one of the most powerful techniques for studying the Fermi surface in metals.

A basis for an understanding of this effect was provided in 1930 by Landau who showed that for a system of free electrons in a magnetic field, the motion of the electrons parallel to the field is classical while the motion perpendicular to the field is quantized. These ideas were shown by Peierls in 1933 to hold for free electrons in a metal (spherical Fermi surface). As a consequence, the free energy F of the system and, hence, the magnetic moment $M = -\partial F / \partial H$ oscillate with magnetic field H . These results were extended by Blackman in 1938 to the case of ellipsoidal energy surfaces, accounting quite well for the experimental results obtained on Bi.

The importance of the de Haas-van Alphen effect as a tool in studying the Fermi surface was perhaps not fully appreciated until Onsager (1952) showed that the frequencies of the oscillations are directly proportional to extremal cross-sectional areas of the Fermi surface perpendicular to the magnetic field. Onsager arrived at this result by applying the Bohr-Sommerfeld quantization condition to the electron orbits normal to the field. If \mathbf{p} is the electronic momentum and if \mathbf{A} is the vector potential, then $\oint \left(\mathbf{p} - \frac{e\mathbf{A}}{c} \right) \cdot d\mathbf{l} = (n + \gamma)h$,

where $\left(\mathbf{p} - \frac{e}{c} \mathbf{A} \right)$ is the canonical momentum, n is an integer, γ is a phase factor and h is Planck's constant. From the Lorentz force equation

$\dot{\mathbf{p}} = \frac{e}{c} (\mathbf{v} \times \mathbf{H})$, it follows that electron paths in momentum space have the same shape as those in real space but changed in scale by a factor $\frac{e}{c} H$ and turned through 90° . We note that $\oint \mathbf{A} \cdot d\mathbf{l} = \int \nabla \times \mathbf{A} \cdot d\mathbf{S} = HS$, where S is the area of the orbit in real space. Furthermore, the area of the orbit in momentum space \mathcal{A}_p is just $\left(\frac{eH}{c} \right)^2$ times the area in real space, so it follows that $\mathcal{A}_p = (n + \gamma) \frac{ehH}{c}$. Since the momentum is related to the wave number k by $p = \hbar k$, the area in k -space is given by $\mathcal{A}_k = (n + \gamma) \frac{2\pi eH}{ch}$.

If one now considers a free electron metal near absolute zero, the surfaces of constant energy ($E \propto k^2$) in k -space will be a quasi-continuous set of spheres up to some maximum size corresponding to the Fermi energy E_F . When a magnetic field is applied in the z direction, these surfaces will degenerate into a discrete set of cylinders, a consequence of the quantization of states in the x - y plane. This is shown schematically in Fig. 1. As indicated above, the cross-sectional area of the n th cylinder will be given by $\mathcal{A}_k(n)$. Each of these permitted states is highly degenerate, but as the field is increased, and the cylinders expand through the Fermi surface, they must give up their electrons to inner cylinders of lower quantum number. For the inner cylinders, this leads to a rather smooth variation in the free energy of the system, but for the outer cylinder, the occupied length decreases very rapidly as the cylinder approaches the extremal cross section of the Fermi surface (\mathcal{A}_0). This rapid depopulation of the n th cylinder at a critical field given by $1/H_n = 2\pi(n + \gamma)e/\hbar c \mathcal{A}_0$ completes an oscillation in the free energy of the system. As the field is increased further, the process is repeated, giving oscillations in the energy and hence the susceptibility

$\chi = -\frac{1}{H} \frac{\partial F}{\partial H}$ which are periodic in $(1/H)$ with period $\Delta(1/H) = 2\pi e/\hbar c \mathcal{A}_0$. The same result occurs for a Fermi surface of more complicated shape, and if the surface consists of several pieces or sections, there will be oscillations related to the extremal cross-sectional area of each section. Observed de Haas-van Alphen periods found in

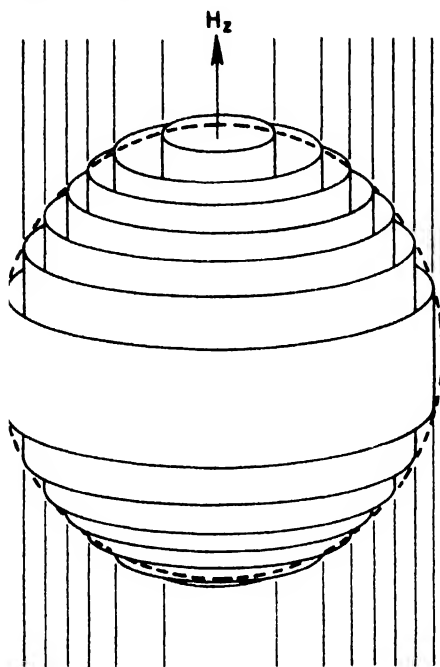


Fig. 1. Quantized orbits in a free electron metal illustrating the origin of the de Haas-van Alphen effect. The dashed sphere denotes the Fermi level.

different metals range in value from $\approx 10^{-5}$ to $\approx 10^{-9}$ gauss¹ corresponding to Fermi surface cross sections from $\approx 10^{-13}$ to $\approx 10^{-17}$ cm². A study of the de Haas-van Alphen effect frequencies (proportional to \mathcal{A}_0) for various directions of the applied field relative to the crystal axes can thus be of great value in mapping out the geometrical shape of the Fermi surface.

At temperatures above absolute zero, the quantized orbits or energy levels (Landau levels) are not sharp but are "smeared out" over an energy range of $\approx kT$ by electron collisions with lattice vibrations. This results in a rapid decrease in the amplitude of the oscillations with increasing temperature. The effect of electron collisions with impurities results in a further reduction in amplitude. A general requirement for the observation of the effect is that $\omega\tau \geq 1$, where ω is the cyclotron frequency, related to the time for an electron to complete an orbit, and τ is the mean time between collisions. For this reason, experiments are generally carried out at liquid-helium temperatures ($\approx 4.2^\circ\text{K}$) on ultrahigh-purity crystals.

A detailed theoretical treatment of the de Haas-van Alphen effect was carried out by Lifshitz and Kosevich.⁴ The result for the free energy can be expressed as

$$F \propto TH^{3/2} \exp\{-2\pi^2 k(T+x)cm^*/ehH\}$$

$$\cos\left\{ch\mathcal{A}_0/eH \mp \frac{\pi}{4} - \delta\right\}$$

for the case $2\pi^2 k(T+x)cm^*/ehH \gg 1$. Here k is the Boltzmann constant, δ is a phase factor, T is

the absolute temperature, and x is a factor which takes into account the impurity collision broadening and is related to the relaxation time of the electron (or hole). The quantity m^* is the "effective mass" or "cyclotron mass" of the electron (or hole) for the extremal cross section and is related to the curvature of the energy surface. It can be seen from the above expression that by determining the temperature and field dependences of the amplitude of the oscillations, one can find the effective mass m^* and the scattering factor x .

Of the various techniques for observing the de Haas-van Alphen effect, the torsion balance method and the pulsed field method have been most common. In the former, the sample is suspended in a uniform transverse field (usually from 0 to 40 kilogauss) and the torque exerted on the sample is recorded as a function of field strength. This method has the advantages of high sensitivity and accuracy, but it cannot detect oscillations arising from spherical Fermi surface segments. Furthermore, available transverse field intensities place limitations on its use for very high de Haas-van Alphen frequencies (corresponding to large Fermi surface areas). In the pulsed field method, the specimen is situated inside of a balanced coil system which detects the changing magnetization in a varying field. Magnetic fields rising to 100 to 200 kilogauss in times on the order of 10 msec are produced through discharge of a large condenser bank into a copper-wound solenoid. The high fields enable one to study very high de Haas-van Alphen frequencies. In addition, the method detects isotropic surface segments.

Another technique which combines some of the advantages of both of the above methods uses a dc magnet modulated by a sinusoidally varying field with amplitude smaller than the period of the de Haas-van Alphen oscillations. A signal proportional to the de Haas-van Alphen effect is detected by means of a small balanced coil system surrounding the sample as in the pulsed field method. This technique is particularly suited for use with superconducting solenoids which produce extremely large, high-stability fields with low power requirements.

A. C. THORSEN

References

1. Schoenberg, D., in Gorter, C. J., Ed., "Progress in Low Temperature Physics," New York, Interscience Publishers, 1957.
2. Chambers, R. G., *Can. J. Phys.*, **34**, 1395 (1956).
3. Pippard, A. B., in Stickland, A. C., Ed., *Rept. Progr. Phys.*, **23**, 176 (1960).
4. Lifshitz, I. M., and Kosevich, A. M., *J. Exp. Theoret. Phys.*, **29**, 730 (1955).

Cross-references: FERMI SURFACE, MAGNETISM.

DENSITY AND SPECIFIC GRAVITY

Density is a fundamental property of matter which is a measure of the compactness of its particles. Density is expressed as the ratio of

mass to volume and depends on the composition of the specimen, its homogeneity, temperature, and especially in the case of gases, on pressure. Density is a property used to identify elements, compounds, and mineral specimens. In the case of minerals where homogeneity varies, a range of densities may be cited as typical of samples of the mineral.

Gases. One way to determine the density of a gas is to weigh an evacuated bulb and then weigh it again filled with the gas, both weighings at known temperature and pressure. The bulb must be dry and free from condensed moisture. The weighing must be corrected for the buoyant force of the air on the bulb. If a two-pan balance is used, an evacuated counterpoise bulb of the same volume may be used. On a single-pan balance, the buoyancy of the air must be calculated for accurate work.

The buoyant effect is the basis for a sensitive method of finding the density of a gas. In the Edwards gas density balance^{3,4} a sealed glass cylinder is at one end of a balance beam and a counterweight at the other. This equipment is mounted within a gastight chamber which is attached to a manometer. Gases are introduced into the chamber—first air and then, after complete removal of the air, the gas whose density is to be measured. The pressure exerted to restore the beam to equilibrium is read in each case. The densities of the two gases are inversely proportional to the pressures required to attain equal buoyant force on the beam within the chamber. A relatively large sample of gas is needed for this method.

The Schilling effusion method for determining the density of a gas⁵ involves measuring the time for a given gas sample of given volume and temperature to escape through a known orifice. This time is compared with the time for an equal volume of dry air under the same conditions. The density of the sample is proportional to the square root of the time of escape.

Liquids. The simplest method of finding the density of a liquid at a specified temperature is to float a hydrometer of suitable range in it and to read the value of the density from the graduations on the hydrometer stem. Such testing is used for the electrolyte in automobile and boat storage batteries and for antifreeze fluids in automobile radiators.

More precise measurement can be made with a Westphal balance. A sinker is weighed in air and again while it is suspended in water. It is then weighed while it is suspended in the liquid to be measured. Since the same volume of liquid is displaced by the sinker in both cases, the buoyant forces are proportional to the densities of the liquids.

A pycnometer is a small glass bottle fitted with a ground-glass stopper through which a capillary hole extends lengthwise. When the bottle is filled with liquid at a given temperature and the contents fill the capillary tube completely (or in some cases to a mark), the pycnometer holds the volume marked on it, or it may be calibrated with a liquid

of known density. The pycnometer must first be weighed dry at the specified temperature, and the weighing must be corrected for the buoyancy of the air. The weight of the contents divided by the volume of the pycnometer is the density of the liquid. Other designs of pycnometers are available, many of them refined for special purposes.

Solids. A wide-mouthed pycnometer allows the introduction of a solid whose density is to be measured. In the Russell form, the stopper has a long graduated tube attached to it and a reservoir above that which can be filled with a dense liquid such as mercury. The apparatus is placed upright and the weight of mercury in the tube is measured. The apparatus is inverted so that the mercury runs into the reservoir. A solid of known weight is placed within the pycnometer, the stopper is inserted, and the liquid is permitted to return. The new and higher reading of the mercury in the stem subtracted from the original reading gives the volume of the solid within the pycnometer. Air bubbles must be avoided.

Mixtures of "density liquids" such as methylene iodide (3.325 g/ml) and benzene (0.87901 g/ml) or similar pairs of miscible liquids are made in such proportions that their densities range between the values given. An insoluble solid has the same density as that of the liquid in which it fails to float or sink. The density of the liquid mixture is checked by the pycnometer or other method.

Both methods depend on the principle of Archimedes that a solid immersed in a liquid has a buoyant force exerted on it proportional to the weight of the volume of the displaced liquid. A simple density measurement can be made by weighing an object in air and then in water. The weight in air divided by the loss of weight in water (density 1 g/ml) is the density of the solid. In case a liquid other than water is used, the volume can be found by dividing the loss of weight of the solid by the density of the liquid in which it is immersed, thus extending the usefulness of the method to solids which are soluble in water but which are insoluble in benzene and other liquids.

A density gradient tube consists of two liquids mixed so that the density gradually and regularly increases from top to bottom. A sample is dropped into the tube, and its density is estimated from the position where it rests within the tube.

Specific Gravity. Specific gravity is the ratio between the density of a specimen and the density of a standard material. For liquids and solids, that standard is ordinarily pure water at 4°C (precisely 3.98°C). Mercury has a density of 13.6 g/ml; water has a density of 1.00 g/ml. Their ratio 13.6 to 1.00 or 13.6, is the specific gravity of mercury. Notice that specific gravity has no units of measurement while density requires measurement units. In the English system, densities are 62.4 times as great as those for solids and liquids in the metric system. Water is 1 g/ml in the metric system, but 62.4 lb/ft³ in the English system. Specific gravities come out numerically the same if English units are used

for the density measurements, as they do for metric units.

For gases, hydrogen (0.08988 g/ml at 760 mm of Hg (torr), and 273°K) and air are the standards. If air is the standard (1.29 g/liter at STP), the specific gravity of hydrogen is 0.0656 and that of carbon dioxide is 1.53.

The specific gravity of sulfuric acid of one concentration is expressed as 1.834.²⁰ This means that sulfuric acid at 20°C is 1.834 times as dense as water at 4°C. The density of this acid is 1.834 g/ml at 20°C.

Density measurements are readily made to four significant figures, and with refinements to seven or eight. Density measurements are sufficiently precise to follow the distribution of isotopes in the course of the electrolysis of lithium fluoride.⁹ The per cent of D₂O in H₂O can be found readily by measuring the density of the mixture.

ELBERT C. WEAVER

References

1. Weissberger, A. "Technique of Organic Chemistry," New York, Interscience Publishers, 1959.
2. Weaver, Elbert C., "Specific Gravity," in Clark and Hawley, Eds., "Encyclopedia of Chemistry," New York, Reinhold Publishing Corp., 1957.
3. Edwards, J. D., *Natl. Bur. Std. Tech. Paper*, **89**, 1917.
4. Smith, F. A., *et al*, *Natl. Bur. Std. Misc. Publ.* **177**, 1947.
5. Kunberger, A. F., (Ch.) "Gas Analysis and Testing of Gaseous Materials," Third edition, p. 321 American Gas Association, 1929.
6. Hutchinson and Johnston, *J. Am. Chem. Soc.*, **62**, 3165 (1940).

DIAMAGNETISM

Magnetic susceptibility is defined as $\chi_m = M/H$, where M is the magnetic moment per gram (gram susceptibility) or per mole (mole susceptibility) that is induced by an external magnetic field strength H . If $M > 0$, the susceptibility is *paramagnetic*; if $M < 0$, the susceptibility is *diamagnetic*. Whereas most magnetic phenomena, including PARAMAGNETISM, are manifestations of ELECTRON SPIN, diamagnetism reflects electron angular momentum.

If an external field strength H is applied to a conductor so as to change the number of lines of flux that thread through it, there is induced in the conductor an electric current whose associated magnetic field opposes the change (*Faraday's Law of Induction and Lenz's Law*). In most conductors the current I that is thus induced is rapidly dissipated as heat through the I^2R loss, where R is the electrical resistance. These currents are known as *eddy currents*, and they are of great practical interest in ac applications. However, such transients do not influence the dc measurement of χ_m . There are three other classes of electron-momentum change induced by H that are not dissipated: electron currents in SUPERCONDUCTORS, where the resistance is $R = 0$;

currents of atomic dimension induced in atoms or molecules or the atomic "core" electrons of solids; and microscopic conduction-electron helical currents having quantized helical radii.

In a superconductor, switching on of an H induces eddy currents that permanently shield the inside of the conductor from penetration by the magnetic-field lines. Therefore the superconductor is an ideal diamagnet, except for a small skin depth at the surface. If a superconductor is cooled through the normal-conducting \rightleftharpoons superconducting transition temperature in the presence of H and after the eddy currents induced in the normal-conducting state have been dissipated, the field lines are rapidly expelled from the superconductor (*Meissner effect*). This proves that the ideally diamagnetic state is thermodynamically stable.

An external field H superposes on the motion of atomic or molecular electrons (or the atomic core electrons in solids) a common circular motion about H of angular frequency ω_L , $eH/2mc$, where e/m is the electronic charge-to-mass ratio (*Larmor's theorem*). This atomic current produces an atomic moment that is proportional to the square of the distance of a classical electron from the nucleus, $r_i \sim 1\text{\AA}$. Therefore the diamagnetic contribution from electrons localized about an atomic nucleus is

$$\chi_m^{\text{core}} = -(Ne^2/6mc^2) \sum_i r_i^2,$$

where N is the number of atoms per gram (or mole).

In addition to macroscopic eddy currents, conduction electrons tend to move in microscopic helical paths in the presence of an H . The contribution to χ_m from this helical motion is a purely quantum mechanical effect. The radii of the H -induced helical paths are quantized, which leads to a "bunching" of the energy levels within an energy band of conducting states, and at large H , these "bunches" can be resolved. They are known as *Landau levels* because Landau¹ first presented the quantum mechanical theory of conduction-electron diamagnetism, which for single parabolic energy bands gives

$$\chi_m^{\text{cond}} = -\frac{1}{3}\mu_B^2 N(E_F),$$

where μ_B is the *Bohr magneton* and $N(E_F)$ is the density of energy levels at the *Fermi energy* E_F . Since $N(E_F)$ oscillates with H as successive Landau levels pass through E_F , χ_m^{cond} shows oscillations in large H (*DE HAAS-VAN ALPHEN EFFECT*). Transitions between Landau levels, which are split by an energy $\hbar\omega_p = eH/m^*c$, may be induced by an electromagnetic field of angular frequency ω_p . This gives rise to resonance power absorption as ω passes through ω_p (*cyclotron or diamagnetic resonance*). These two effects are used to map out the contours in momentum space of the Fermi energies in metals.

JOHN B. GOODENOUGH

Reference

1. Landau, L. D., *Z. Physik* **64**, 629 (1930).

DIELECTRIC THEORY

A dielectric is a material having electrical conductivity low in comparison to that of a metal. It is characterized by its dielectric constant and dielectric loss, both of which are functions of frequency and temperature. The dielectric constant is the ratio of the strength of an electric field in a vacuum to that in the dielectric for the same distribution of charge. It may also be defined and measured as the ratio of the capacitance C of an electrical condenser filled with the dielectric to the capacitance C_0 of the evacuated condenser:

$$\epsilon = C/C_0$$

The increase in the capacitance of the condenser is due to the polarization of the dielectric material by the applied electric field. The terms "specific inductive capacity" or "permittivity" are occasionally used instead of dielectric constant. The constant ϵ appearing in the Coulomb law of force is called the permittivity, but it is also commonly called the dielectric constant. The relative permittivity or dielectric constant is the ratio ϵ/ϵ_0 , where ϵ_0 is the permittivity or dielectric constant of free space. In the mks system of units, the dielectric constant of free space is 8.854×10^{-12} farad/m, while in the esu system the relative and the absolute dielectric constants are the same. The relative dielectric constant, which is dimensionless, is the one commonly used. When variation of the dielectric constant with frequency may occur, the symbol is commonly primed. When a condenser is charged with an alternating current, loss may occur because of dissipation of part of the energy as heat. In vector notation, the angle δ between the vector for the amplitude of the charging current and that for the amplitude of the total current is the loss angle, and the loss tangent, or dissipation factor, is

$$\tan \delta = \frac{\text{Loss current}}{\text{Charging current}} = \frac{\epsilon''}{\epsilon'}$$

where ϵ'' is the loss factor, or dielectric loss, of the dielectric in the condenser and ϵ' is the measured dielectric constant of the material.

At low frequencies of the alternating field, the dielectric loss is normally zero and ϵ' is indistinguishable from the dielectric constant ϵ_{dc} measured with a static field. Debye has shown that

$$\frac{\epsilon_{dc} - 1}{\epsilon_{dc} + 2} = \frac{4\pi N_1}{3} \left(\alpha_0 + \frac{\mu^2}{3kT} \right) \quad (1)$$

where N_1 is the number of molecules or ions per cubic centimeter; α_0 is the molecular or ionic polarizability, i.e., the dipole moment induced per molecule or ion by unit electric field (1 esu = 300 volts/cm); μ is the permanent dipole moment possessed by the molecule; k is the molecular gas constant, 1.38×10^{-16} , and T is the absolute temperature. An electric dipole is a pair of electric charges, equal in size, opposite in sign, and very close together. The dipole moment is the product of one of the two charges by the distance between them.

In Eq. (1) $\mu^2/3kT$ is the average component in the direction of the field of the permanent dipole moment of the molecule. In order that this average contribution should exist, the molecules must be able to rotate into equilibrium with the field. When the frequency of the alternating electric field used in the measurement is so high that dipolar molecules cannot respond to it, the second term on the right of the above equation decreases to zero and we have what may be termed the optical dielectric constant ϵ_∞ , defined by the expression

$$\frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} = \frac{4\pi N_1}{3} \alpha_0 \quad (2)$$

ϵ_∞ differs from n^2 , the square of the optical refractive index for visible light, only by the small amount due to infrared absorption and to the small dependence of n on frequency, as given by dispersion formulas. It is usually not a bad approximation to use $\epsilon_\infty = n^2$. The general Maxwell relation $\epsilon' = n^2$ holds when ϵ' and n are measured at the same frequency. The Debye equation may be written in the form

$$\frac{\epsilon_{dc} - 1}{\epsilon_{dc} + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} = \frac{4\pi N_1}{9kT} \mu^2 \quad (3)$$

A much better representation of the dielectric behavior of polar liquids is given by the Onsager equation

$$\frac{\epsilon_{dc} - 1}{\epsilon_{dc} + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} = \frac{3\epsilon_{dc}(\epsilon_\infty + 2)}{(2\epsilon_{dc} + \epsilon_\infty)(\epsilon_{dc} + 2)} \frac{4\pi N_1 \mu^2}{9kT} \quad (4)$$

Anomalous dielectric dispersion occurs when the frequency of the field is so high that the molecules do not have time to attain equilibrium with it. One may then use a complex dielectric constant

$$\epsilon^* = \epsilon' - j\epsilon'' \quad (5)$$

where $j = \sqrt{-1}$. Debye's theory of dielectric behavior gives

$$\epsilon^* = \epsilon_\infty + \frac{\epsilon_{dc} - \epsilon_\infty}{1 + j\omega\tau} \quad (6)$$

where ω is the angular frequency (2π times the number of cycles per second) and τ is the dielectric relaxation time. Dielectric relaxation is the decay with time of the polarization when the applied field is removed. The relaxation time is the time in which the polarization is reduced to $1/e$ times its value at the instant the field is removed, e being the natural logarithmic base.

Combination of the two equations for the complex dielectric constant and separation of real and imaginary parts gives

$$\epsilon' = \epsilon_\infty + \frac{\epsilon_{dc} - \epsilon_\infty}{1 + \omega^2\tau^2} \quad (7)$$

$$\epsilon'' = \frac{(\epsilon_{dc} - \epsilon_\infty)\omega\tau}{1 + \omega^2\tau^2} \quad (8)$$

These equations require that the dielectric constant decrease from the static to the optical dielectric constant with increasing frequency, while the dielectric loss changes from zero to a maximum value ϵ''_m and back to zero. These changes are the phenomenon of anomalous dielectric dispersion. From the above equations, it follows that

$$\epsilon_m'' = (\epsilon_{dc} - \epsilon_\infty)/2 \quad (9)$$

and that the corresponding values of ω and ϵ' are

$$\omega_m = 1/\tau \quad (10)$$

and

$$\epsilon_m' = (\epsilon_{dc} + \epsilon_\infty)/2 \quad (11)$$

The symmetrical loss-frequency curve predicted by this simple theory is commonly observed for simple substances, but its maximum is usually lower and broader because of the existence of more than one relaxation time. Various functions have been proposed to represent the distribution of relaxation times. A convenient representation of dielectric behavior is obtained, according to the method of Cole and Cole, by writing the complex dielectric constant as

$$\epsilon^* = \epsilon' + \frac{\epsilon_{dc} - \epsilon_\infty}{1 + (j\omega\tau_0)^{1-\alpha}} \quad (12)$$

where τ_0 is the most probable relaxation time and α is an empirical constant with a value between 0 and 1, usually less than 0.2. When the values of ϵ'' are plotted as ordinates against those of ϵ' as abscissas, a semicircular arc is obtained intersecting the abscissa axis at $\epsilon' = \epsilon_\infty$ and $\epsilon' = \epsilon_{dc}$. The center of the circle of which this arc is a part lies below the abscissa axis, and the diameter of the circle drawn through the center from the intersection at ϵ_∞ makes an angle $\alpha\pi/2$ with the abscissa axis. When α is zero, the diameter lies in the abscissa axis, there is but one relaxation time, and the behavior of the material conforms to the simple Debye theory. When, as may arise from intramolecular rotation, a substance has more than one relaxation mechanism, or, when the material is a mixture, the observed loss-frequency curve is the resultant of two or more different curves and, therefore, departs from the simple Debye or Cole-Cole curve.

If the dielectric material is not a perfect dielectric, and has a specific dc conductance k' (ohms⁻¹cm⁻¹), there is an additional "dielectric loss

$$\epsilon_{dc}'' = \frac{3.6 \times 10^{12} \pi k'}{\omega} \quad (13)$$

The effective specific conductance is given by

$$k' = \frac{1}{4\pi} \frac{(\epsilon_{dc} - \epsilon_\infty)\omega^2\tau}{1 + \omega^2\tau^2} \quad (14)$$

It is evident from this equation that k' increases with ω , approaching a limiting value, k_∞ , the infinite-frequency conductivity, which is attained

when 1 can be neglected in comparison with $\omega^2\tau^2$, so that

$$k_\infty = \frac{\epsilon_{dc} - \epsilon_\infty}{4\pi\tau} \quad (15)$$

In a heterogeneous material, interfacial polarization may arise from the accumulation of charge at the interfaces between phases. This occurs only when two phases differ considerably from each other in dielectric constant and conductivity. It is usually observed only at very low frequencies, but, if one phase has a much higher conductivity than the other, the effect may increase the measured dielectric constant and loss at frequencies as high as those of the radio region. This so-called Maxwell-Wagner effect depends on the form and distribution of the phases as well as upon their real dielectric constants and conductances. Each type of form and distribution requires special treatment. For a commercial rubber, for example, the observed loss may be

$$\epsilon''(\text{observed}) = \epsilon_{dc}'' + \epsilon''(\text{Maxwell-Wagner}) + \epsilon''(\text{Debye}) \quad (16)$$

CHARLES P. SMYTH

References

- Böttcher, C. J. F., "Theory of Electric Polarization," New York, Elsevier, 1952.
- Debye, P., "Polar Molecules," reprinted by Dover, New York, 1945.
- Frohlich, H., "Theory of Dielectrics," Second edition, London, Oxford University Press, 1958.
- Smyth, C. P., "Dielectric Behavior and Structure," New York, McGraw-Hill Book Co., 1955.

Cross-references: DIPOLE MOMENTS, REFRACTION, RELAXATION.

DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS

According to the principle of Huygens (1629–1695) each point in the space which is touched by a wave gives rise to a spherical secondary wave, which again produces tertiary waves, and so on. Every wave interferes with the next one and quite generally gives rise to diffraction phenomena (see OPTICS, PHYSICAL). Such phenomena in the case of visible light first were observed by F. M. Grimaldi (1618–1663) and mathematically explained by J. Fresnel (1788–1827). G. R. Kirchhoff (1824–1887) gave the first exact mathematical solution of the scalar wave differential equation in terms of a boundary integral. If both the primary and the diffracted rays are parallel (e.g., small source, large distances between diffracting sample and source and detector), we have the experimental conditions of Fraunhofer (1787–1826). With two lenses L_1 and L_2 , a collimator pinhole P in the focal plane of L_1 , and a photographic plate F in the focal plane of L_2 , this condition is fulfilled even for short distances (Fig. 1). Monochromatic light is produced by a Hg lamp with

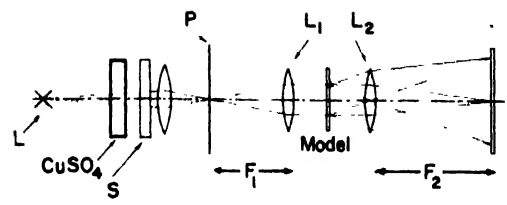


FIG. 1. Equipment for Fraunhofer diffraction.

a Schott filter *S* and an aqueous solution of CuSO4. If *s*₀ and *s* are unit vectors in the direction of the primary and the diffracted beam, and *λ* is the wavelength of the source, then the diffracted intensity *I* is proportional to

$I(\mathbf{b}) \cong f_e^2 f_\theta^2 |R|^2; R(\mathbf{b}) = F(\rho)$ (1)

where

$\mathbf{b} = \frac{\mathbf{s} - \mathbf{s}_0}{\lambda}$ (2)

and

$F = \int e^{-i\pi(\mathbf{b}\cdot\mathbf{x})} dV_x$ (3)

is the symbol of the Fourier transform. *x* is a vector in physical space, *ρ*²(*x*) is the transparency of the object *M* at the end point of the vector *x*, which lies in the plane of the object, *dv_x* is a surface element of the object, *f*_θ² and *f*_{*e*}² are explained in Table 1 and Eq. (2) below Table 1. Figure 2(a) shows an object *ρ*(*x*) in the form of a parallelogram with the edge vectors *L*₁ and *L*₂, and Fig. 2(b) give its Fraunhofer pattern:

$|R|^2 = |S|^2;$

$S(\mathbf{b}) = |\mathbf{L}_1 / \mathbf{L}_2| \frac{\sin \pi(\mathbf{b}\mathbf{L}_1)}{\pi(\mathbf{b}\mathbf{L}_1)} \cdot \frac{\sin \pi(\mathbf{b}\mathbf{L}_2)}{\pi(\mathbf{b}\mathbf{L}_2)}$ (4)

Fraunhofer used assemblies of *N* parallel oriented metal wires with an intermediate distance *d* and the distance *a* (from center to center), in the direction *s*₁. Hence

$\rho(x_1) = \begin{cases} 1 & \text{for } na - \frac{d}{2} \leq x_1 \leq na + \frac{d}{2} \\ 0 & \text{for all other } x_1 \end{cases}$ (5)

TABLE 1. FACTORS *f_e*² AND *f_θ*² FOR DIFFERENT RADIATIONS

		<i>f_e</i> ²	<i>f_θ</i> ²	
1	Visible light	1/λ ²	$\frac{1 + \cos^2 2\theta}{2}$	2θ scattering angle
2	X-rays	$\left(\frac{e^2}{m_0 c^2}\right)^2$	$\frac{1 + \cos^2 2\theta}{2}$	<i>e</i> electric charge of an electron <i>m</i> ₀ rest mass of an electron
3	Electrons	$\frac{m_0 e^2 \lambda^2}{2h^2}$	1/sin ⁴ θ	<i>c</i> velocity of light <i>h</i> Planck's constant
4	Neutrons	Cross section	Polarization factor	λ (de Broglie) wave length

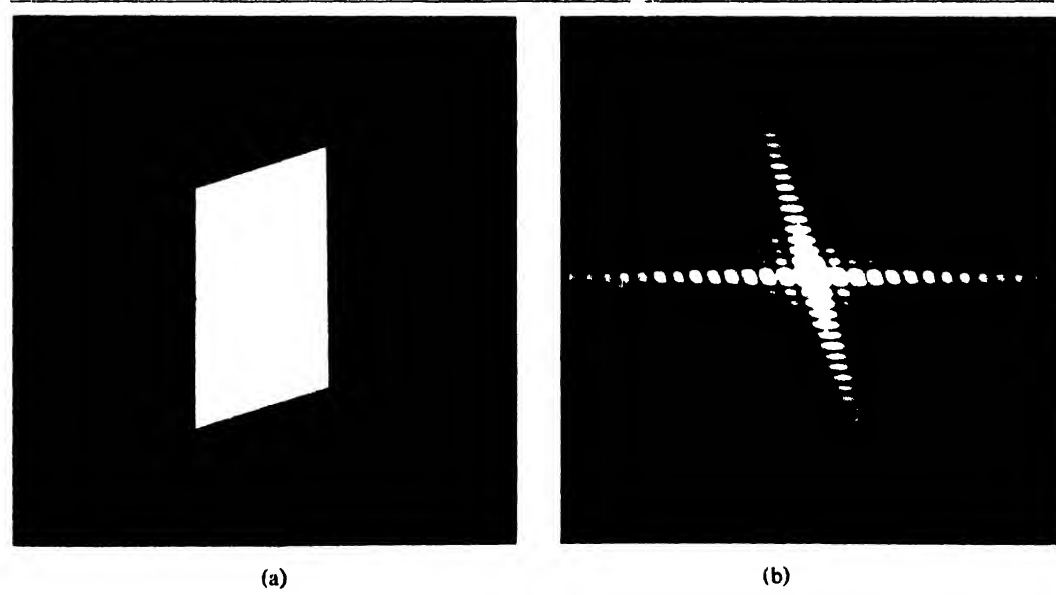


FIG. 2. (a) Parallelogram as diaphragm. (b) Fraunhofer pattern. Secondary maxima of the shape factor *S*² (Eq. (4)).

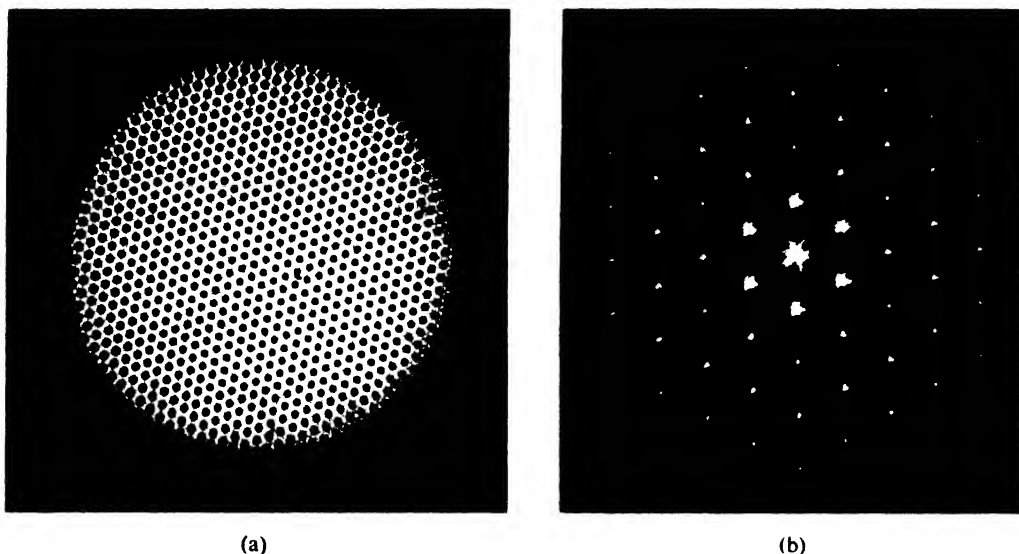


FIG. 3. (a) Steel balls in a crystalline lattice. (b) Bragg reflections with shape factor S^2 .

and

$$I(b_1) = \frac{1}{a} f^2 Z |S|^2 \quad (6)$$

$$f = \frac{\sin \pi b_1 d}{\pi b_1}; \quad Z = \frac{1}{a} \sum_n P(b_1 - n/a);$$

$$S = \frac{\sin \pi b_1 Na}{\pi b_1} \quad (7)$$

f is the Fourier transform of a single slit, S that of the shape of the whole lattice (length Na). Z is the "reciprocal" lattice point function (lattice factor) of the centers of the wire, since $P(b_1 - 0)$ is a normalized point function at $b_1 = 0$. The symbol of convolution \sim is defined for both functions $G_1(\mathbf{b})$ and $G_2(\mathbf{b})$ in Fourier space, and $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ in physical space by

$$\widehat{G_1(\mathbf{b})G_2(\mathbf{b})} = \int G_1(\mathbf{c})G_2(\mathbf{b} - \mathbf{c})d\mathbf{v}_c \quad (8)$$

$$\widehat{g_1(\mathbf{x})g_2(\mathbf{x})} = \int g_1(\mathbf{y})g_2(\mathbf{x} - \mathbf{y})d\mathbf{v}_y \quad (9)$$

In the one-dimensional case of Eq. (6) c is to be replaced by the scalar quantity c_1 and $d\mathbf{v}_c$ by dc_1 . Two-dimensional gratings are of high interest, since their Fraunhofer pattern gives all the information we need to understand the more complicated structural theories of three-dimensional matter. This will be shown below.

The technique of preparing such models is quite easy: In the examples of Figs. 4, 5, 6, 7, 8, 10 and 12, the objects were painted with india ink on paper 15×15 inches and then photographed on fine-grained films 0.5×0.5 inch with steep gradation characteristics (for instance Peruline film F 10). The black ink points now become transparent on a black background. In the models of Figs. 3, 9 and 11, steel balls of 2 to 3 mm diameter were placed into the focal plane of a lens system and photographed with a linear reduction 1:15 on the same film material mentioned above.

By defocusing the system, one obtains diaphragms, where the single balls do not touch each other (Figs. 3 and 9, but not Fig. 11). This is quite advantageous, since the width of the atom-factor f^2 [see Eq. (6)] is larger and more reflections are visible.

Fig. 3(a) shows a model of a two-dimensional ideal periodic "point lattice," and Fig. 3(b) represents its Fourier transform.

$$\rho(\mathbf{x}) = \sum_r P(\mathbf{x} - \mathbf{x}_r); \quad \mathbf{x}_r = p_1 \mathbf{a}_1 + p_2 \mathbf{a}_2 \quad (10)$$

$$R^2 \sim Z(\mathbf{b}) = \frac{1}{|\mathbf{a}_1 \wedge \mathbf{a}_2|} \sum_h P(\mathbf{b} - \mathbf{b}_h); \quad \mathbf{b}_h = h_1 \mathbf{A}_1 + h_2 \mathbf{A}_2 \quad (11)$$

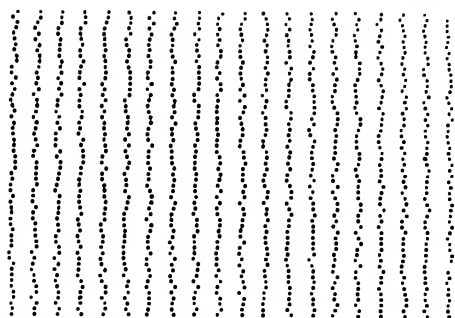
where \mathbf{a}_1 and \mathbf{a}_2 are the vectors of a lattice cell of the model, and \mathbf{A}_1 and \mathbf{A}_2 are those of the "reciprocal" lattice cell (in \mathbf{b} -space)

$$\mathbf{A}_1 = \frac{\mathbf{a}_2 \wedge \mathbf{a}_3}{\mathbf{a}_1(\mathbf{a}_2 \wedge \mathbf{a}_3)}; \quad \mathbf{A}_2 = \frac{\mathbf{a}_3 \wedge \mathbf{a}_1}{\mathbf{a}_2(\mathbf{a}_3 \wedge \mathbf{a}_1)} \quad (12)$$

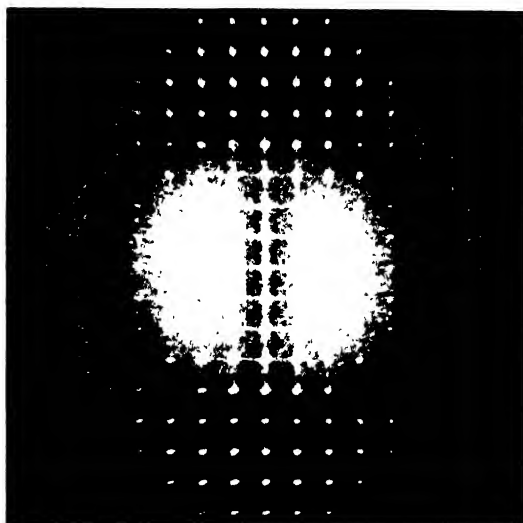
\mathbf{a}_3 is an arbitrary vector orthogonal to \mathbf{a}_1 and \mathbf{a}_2 , and p_1, p_2, h_1 and h_2 are integers. Since the lattice of Fig. 3(a) is bounded if one multiplies $\rho(\mathbf{x})$ by the shape function $s(\mathbf{x})$ of the lattice (which is 1 inside the lattice, and zero out of it) the p -summation can be extended to infinity. From the convolution theorem one obtains generally

$$Fg_1g_2 = \widehat{G_1G_2}; \quad F\widehat{g_1g_2} = G_1G_2 \quad (13)$$

Hence the Fourier transform of the bounded lattice $\rho(\mathbf{x}) \cdot s(\mathbf{x})$ is given again by the convolution product of Eq. (6), where $S(\mathbf{b})$ is the Fourier transform of $s(\mathbf{x})$ and $Z(\mathbf{b})$ is the Fourier transform of Eq. (11). $f^2 = 1$ applies only for "point-like" atoms, otherwise f is the Fourier transform of the shape of each "point."



(a)



(b)

FIG. 4. (a) Linear thermal oscillations without correlations. (b) Debye factor and thermodiffuse background.

M. v. Laue in 1911 prepared an article "Wellenoptik" for the "Encyclopedie der Math. Wissenschaften" and found that Eq. (11) can be easily developed to the Fourier transform of three-dimensional point lattice, where \mathbf{a}_3 is a third lattice vector, non-coplanar to \mathbf{a}_1 and \mathbf{a}_2 . Then \mathbf{b}_h in Eq. (11) must be replaced by

$$\mathbf{b}_h = h_1\mathbf{a}_1 + h_2\mathbf{a}_2 + h_3\mathbf{a}_3; \mathbf{a}_3 = \frac{\mathbf{a}_1 \wedge \mathbf{a}_2}{\mathbf{a}_3(\mathbf{a}_1 \wedge \mathbf{a}_2)} \quad (14)$$

Together with W. Friedrich and P. Knipping using x-rays of a wavelength of the same order of magnitude as that of the atomic distances, M. v. Laue (1912) found three-dimensional diffraction effects in single crystals. $f(\mathbf{b})$ then is the atom form amplitude

$$f(\mathbf{b}) = F(\rho_0) \quad (15)$$

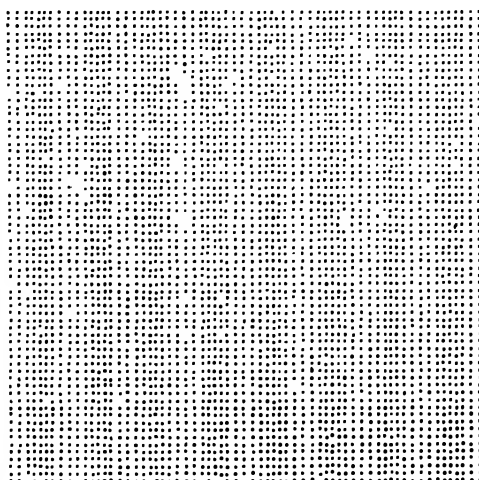
where $\rho_0(\mathbf{x})$ is the electron density distribution of one atom, whose center lies at $\mathbf{x} = 0$.

C. J. Davisson and L. H. Germer (1927) observed the same diffraction phenomena using

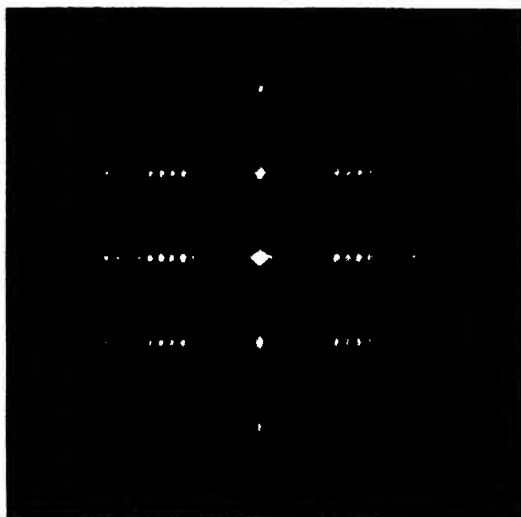
electron beams. $\rho(\mathbf{x})$ must then be understood as density distribution of both the electrons (negative) and protons (positive). If single diffraction processes occur, Eq. (6) remains unchanged.

D. P. Mitchell, P. N. Powers, H. v. Halban and P. Preiswerk (1936) found the same diffraction phenomena using thermal neutrons. In this case $\rho(\mathbf{x})$ is the density distribution of the nuclei, each one weighted by the mean of the square root of its respective cross section. The proportional factors f_e^2 , f_n^2 of Eq. (1) for the different radiations are given in Table 1.

The vector \mathbf{b} defined by the integral of Eq. (3) expands the three-dimensional Fourier space and

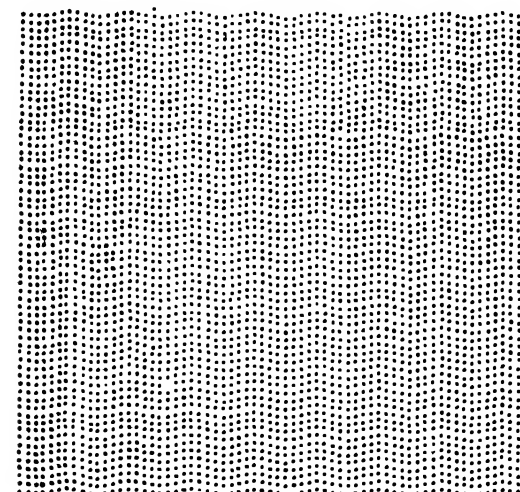


(a)

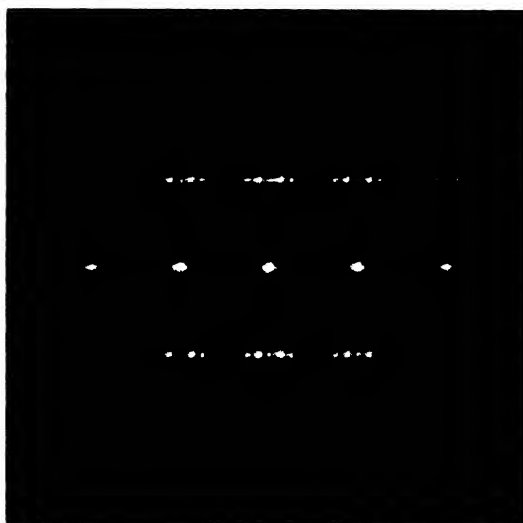


(b)

FIG. 5. (a) A single longitudinal wave. (b) Extra Laue spots.



(a)



(b)

FIG. 6 (a) A single transversal wave. (b) Extra Laue spots.

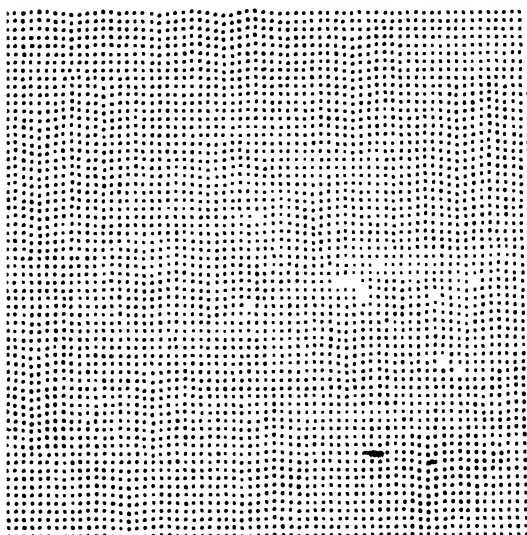
is connected with the unit vectors s , s_0 of the diffracted and primary beam by Eq. (2). This is the construction of Ewald (1914). For a fixed λ and s_0 all values $R(b)$ are in reflection position, which lie on a sphere with radius $1/\lambda$ and the center at $b_0 = s_0/\lambda$.

P. Debye (1915) found that the molecules in the gaseous state give rise to diffraction patterns depending on the structure of the single molecules, without any intermolecular interferences.

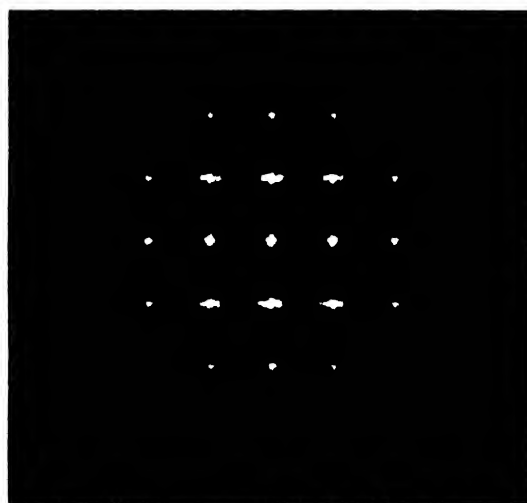
In 1927, F. Zernike and I. A. Prins discussed quantitatively diffraction phenomena of liquids and laid down the fundamentals of the structure analysis of amorphous matter. Moreover, P. Debye (1913) and I. Waller (1927) found that real crystals never have a periodic lattice similar to that of Fig. 3(a), since the atoms show thermal

oscillations around their ideal positions. All these different phenomena can be studied quite easily with the help of two-dimensional statistical models and their Fraunhofer patterns, since Eq. (1) holds for all diffraction phenomena. In Fig. 4(a) we have the "frozen" structure of a point lattice with thermal oscillations. They are quite anisotropic and occur only in the horizontal direction. If $H(x)$ is the frequency of the center of an atom being at the distance x from its ideal position and if the atoms oscillate independently from each other, then R^2 is given by

$$R^2(b) = Nf^2(1 - D^2) + \frac{1}{v_r} f^2 D^2 Z |S|^2; \quad D(b) = F(H) \quad (16)$$

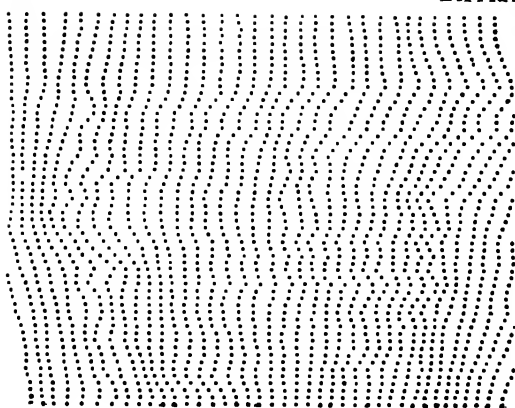


(a)



(b)

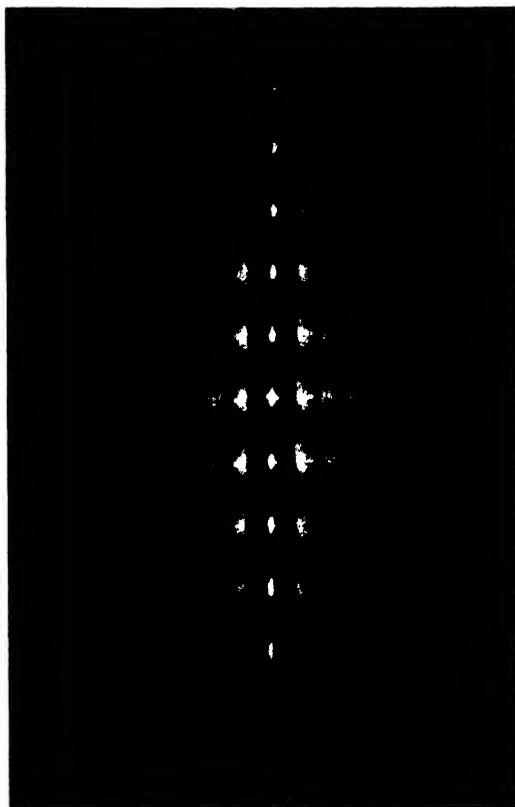
FIG. 7. (a) Twelve damped transversal waves. (b) Diffuse extra Laue spots.



(a)



(c)



(b)

FIG. 8. (a) Monoparacrystal with linear horizontal coordination statistics. (b) Its Fraunhofer pattern. The reflections (o , h_2) are crystalline, all others are more or less diffuse (c) Small-angle x-ray pattern from β -feather-keratin (Bear-Rugo, 1951).

(v_r = volume of a lattice cell). In Fig. 4(b) can be clearly recognized the first term of Eq. (16) as a diffuse background. It has a structure in the horizontal direction. The "Bragg reflections" are weakened by the "Debye-Waller factor" D^2 , the more, the stronger is the diffuse background ($1 - D^2$). In nature there exist correlations between the different oscillations and "elastic waves" with an "acoustical" and "optical" frequency creep through the lattice.

Fig. 4(a) gives a single undamped longitudinal and sinusoidal wave with a horizontal wave vector b_x and a wavelength λ_x and amplitude a_x

$$\lambda_x = \frac{1}{|b_x|} = 8a_x; |a_x| = \frac{1}{4}a \quad (17)$$

According to the theories of M. v. Laue (1927) and Laval (1941) in Fig. 5(b) at the reciprocal lattice points

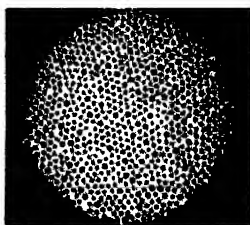
$$\mathbf{b} = \mathbf{b}_h + m\mathbf{b}_x \quad (18)$$

"Extra Laue spots" ("Laval spots") occur. According to conventional theories, $m=1$ is called one-phonon scattering, $m=2$ is two-phonon scattering and so on; and it is said that the frequency ν_0 of the incident radiation is changed into $\nu_0 + m\nu_x$ ("inelastic scattering", ν_x -frequency of the phonons).

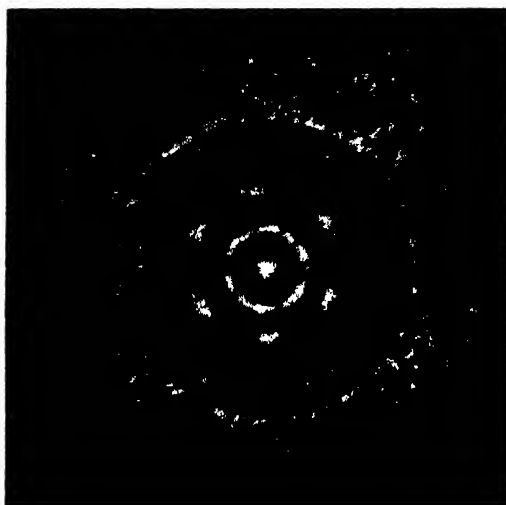
In Fig. 6(a) we have again a horizontal wave with the same wave vector [Eq. (17)], which now is transverse. From Laue's theory it follows, that the intensity of a Laue spot of the m th order is proportional to the square of the Bessel function

$$I_m(2\pi(\mathbf{b}\mathbf{a}_x))$$

where m is the order of the Bessel function and \mathbf{a}_x is the amplitude vector of the elastic wave. Hence in Fig. 6(b), strong Laval spots occur only in the vertical \mathbf{b}_h direction; in Fig. 5(b), in the horizontal \mathbf{b}_h direction. Since $I_1(I_2)$ has its maximum at $\mathbf{b}\mathbf{a}_x = 0.25$ (0.5), in Figs. 5(b) and 6(b), the Laval spot $m=1$ at the reflection h_1 ,



(a)



(b)

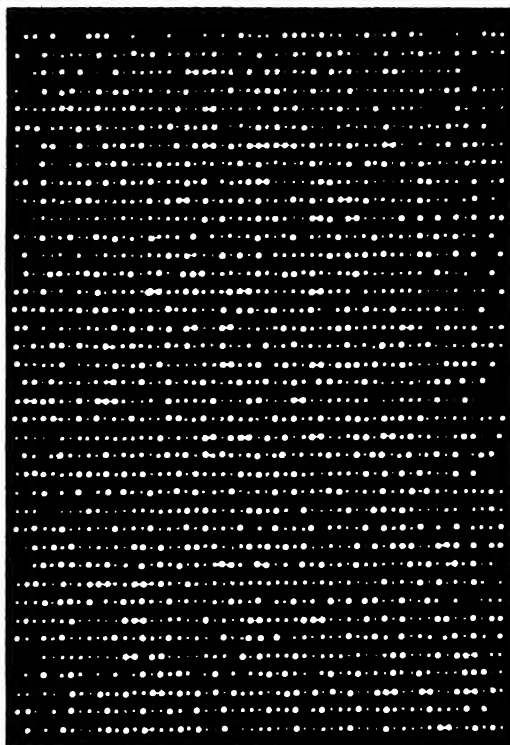
FIG. 9. (a) Mixed single paracrystal (steel balls). (b) Intensity function (Eq. (6) and (19)).

$h_2 = 1.0$ (0.1) is much stronger and in the 2.0 (0.2) reflection, much weaker, than the Laval spot $m = 2$.

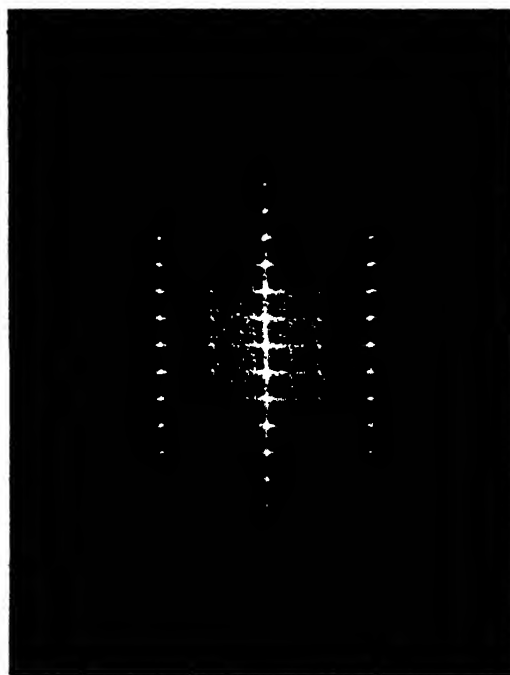
In nature, one never finds such sharp Laval spots, but a completely smooth "thermodiffuse" scattering, and it is said that a "white" spectrum of undamped elastic waves exists.

In Fig. 7(a) strongly damped transverse waves according to Eq. (17) consisting only of two maxima and minima are introduced. Now in Fig. 7(b) quite diffuse spots appear. Hence, in nature, such damped waves could also occur, and if they had a spectral distribution, they could give rise to the same observable thermodiffuse scattering. According to Debye's theory of heat capacity, every wave has a different amplitude a_i , following a Boltzmann statistic. Then in certain regions of a single crystal, the atoms oscillate statistically with different amplitudes than in others. As a result, the lattice cells exhibit different sizes and paracrystalline distortions occur.

Close to the melting point, the amplitudes a_i of the elastic waves become so large that the electron clouds of the atoms suffer large deformations, which damp these waves more and more.

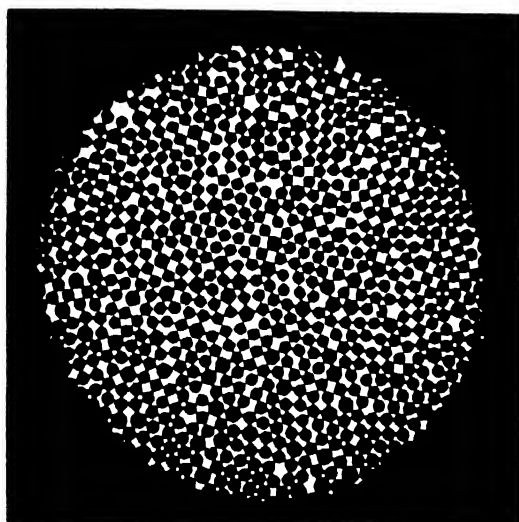


(a)

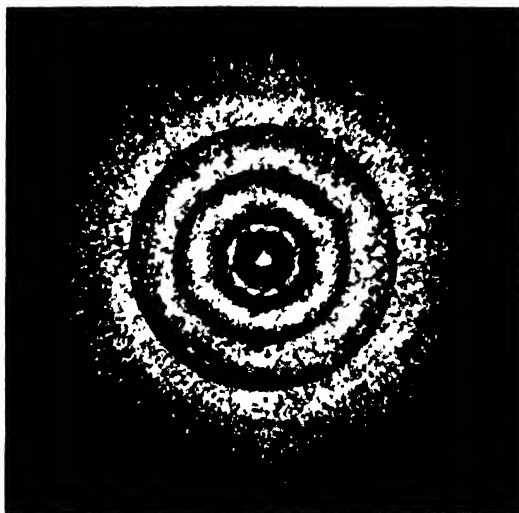


(b)

FIG. 10. (a) Mixed single crystal with "cooperative forces". (b) Background with diffuse walls (*Nahordnung*).



(a)

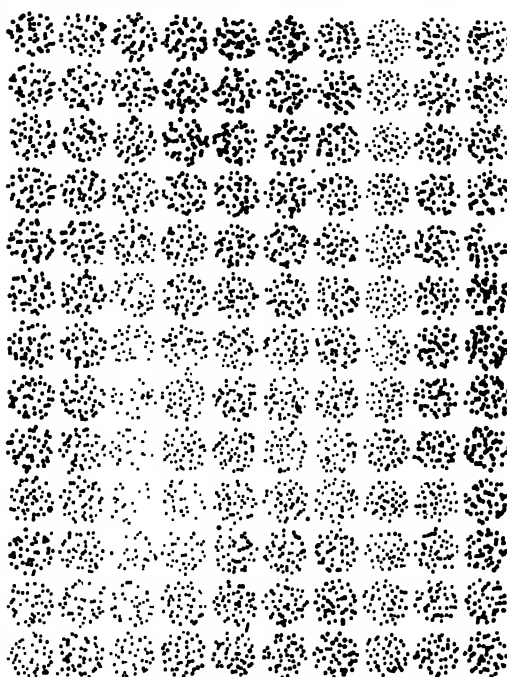


(b)

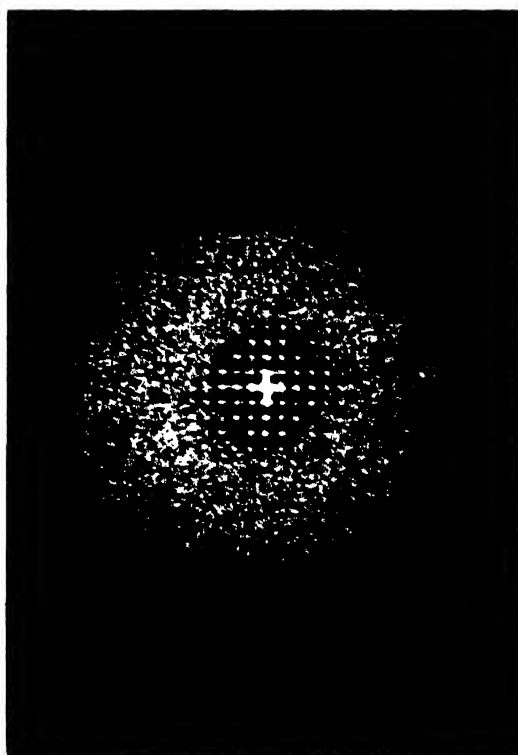
FIG. 11. (a) Polymicroparacrystalline assembly of steel balls. (b) Intensity function of "amorphous" matter.

Another physical reason for paracrystalline distortions below the melting point exists if the "motives" in the single lattice cells have different shapes. Such phenomena have been observed in the "macrolattices" of natural and synthetic high polymers.

Figure 8(a) shows a paracrystalline model, where both coordination statistics H_k are horizontal line functions. Figure 8(b) represents its Fraunhofer pattern and Fig. 8(c) the x-ray small-angle pattern of the β -keratin of the quill of a sea gull (Bear and Rugo (1951)). The cell edge a_2 in the vertical direction parallel to the fiber axis has a constant length of 185Å but statistically



(a)



(b)

FIG. 12. (a) Zirconium crystal with discrete electrons. (b) Rayleigh scattering of the electrons.

changes its direction (with respect to the macroscopic fiber axis) within $\pm 5^\circ$, while the orthogonal edge length a_1 has an average value of 34\AA and changes its length statistically within $\pm 2.5\text{\AA}$. As a result of the van der Waals forces, which allow a variation of the length a_1 orthogonal to the fiber axis, the homopolar forces along a_2 have freedom to change their direction of the molecular chain, within one paracrystal. This paracrystal itself, therefore, shows a flexible character in atomic dimensions.

Equation (6) holds again, if one replaces the crystalline lattice factor Z of Eqs. (7) and (11) by the paracrystalline lattice factor

$$Z(b) = \frac{1}{v} \prod_{k=1}^3 \text{Re} \frac{1 + F_k(b)}{1 - F_k(b)} \quad (19)$$

$$F_k(b) = F(H_k) \quad (20)$$

H_k is a so-called coordination statistic. $H_k(\mathbf{x})$ is the "a priori" probability of finding a cell edge vector $\mathbf{a}_k = \mathbf{x}$ in a certain paracrystalline lattice cell.

For the same reason, in mixed crystals consisting of atoms of different sizes, paracrystalline distortions can also occur.

Figure 3(a) showed a model of steel balls of the same size, building up a crystalline lattice. Figure 3(b) represents its Fourier transform. In Fig. 9(a) steel balls of different sizes built up a paracrystalline lattice, whose Fraunhofer pattern [Fig. 9(b)] exhibits characteristic features of a paracrystalline lattice [cf. Fig. 8(b)].

In Fig. 10(a) another model of a mixed crystal is given without paracrystalline lattice distortions. Now the two kinds of atoms are not distributed totally randomly to the lattice points. Hence the Fraunhofer pattern of Fig. 10(b) shows a diffuse background, which is not given by

$$N(f^2 - f^2) \quad (21)$$

and which presents "diffuse walls" in the vertical direction at $h_1 = \pm \frac{1}{2}$ of a width $\delta h_1 \sim \lambda/4$. This means that in the horizontal direction, rows of about 8 atoms show a kind of "superstructure" (*Nahordnung*, "cooperative order"). This "superstructure" has a psychological background: The technician, painting the models row by row in the horizontal direction was anxious to choose thick and thin atoms quite arbitrarily. However, he did not use a Monte Carlo Method, but tried as arbitrarily as possible to draw statistical sequences. Unfortunately, after he painted a thick atom he tended to choose a thin one, etc. Unconsciously, he introduced "cooperative forces."

Similar diffuse walls are observed in ferroelectric NaNO_2 above the Curie point: In this case, rows of about six NaNO_2 molecules have parallel oriented polar axes, and have built up microferroelectric domains in a statistically paraelectric matrix (M. Canut and R. Hosemann, 1964).

In Fig. 11(a), similar to Fig. 9(a), steel balls of different sizes which now built up a structure of

single small paracrystallites were used. The Fraunhofer pattern, Fig. 11(b), shows the typical features of a liquid or melt or "amorphous" solid. Hosemann and Lemm (1964) have proved, that in molten gold and lead such paracrystals can be found with average diameters of 12\AA to 40\AA .

After having completed the step from crystals to amorphous matter, Fig. 12(a) gives an example of a special gas and Fig. 12(b) at larger b -values, the gas-interferences. Lord Rayleigh (1842-1919) proved that here phase relations between the single scattering centers are destroyed in the average as a consequence of their irregular positions. Since in Fig. 12 only 5200 centers were used, "ghosts" remain in the Fraunhofer pattern. Besides this fluctuation the intensity at large b -values is given by the shape of the single points. Moreover in Fig. 12(a) every 40 points cluster together. The clusters are arranged in a crystalline lattice. Hence, at small angles in Fig. 12(b), "Bragg-reflections" whose intensity is proportional to the squared transform Eq. (15) do occur. ρ_0 is now a forty-point function different for each cluster, and in Eq. (6), f must be replaced by the average \bar{f} . The diffuse background of Fig. 12(b), which now is the statistical fluctuation of the density distribution ρ_0 is given again by Eq. (21). If we replace the word "cluster" by zircon atom and "point" by electron, Fig. 12(a) gives an instantaneous picture of the electron configuration in a zirconium crystal. The Bragg reflections give in this case information about Schrodinger's wave functions for electrons

$$\overline{\rho_0(\mathbf{x})} = \psi\psi^*$$

and the fluctuation term of Eq. (21) gives some information about the structure of a single electron. In reality, Compton processes disturb the diffraction. There is some hope for further detailed studies.

ROLF HOSEMANN

Cross-references: ABERRATION, DIFFRACTION THEORY OF; ELECTRON DIFFRACTION; NEUTRON DIFFRACTION; OPTICS, PHYSICAL; SCHRODINGER EQUATION; X-RAY DIFFRACTION; CRYSTALLOGRAPHY; POLYMER PHYSICS.

DIFFUSION IN LIQUIDS

Diffusion, in a macroscopic sense, is a universal process that leads to the elimination of concentration gradients in gases, solids, or liquids. At the molecular level, it arises because atoms or molecules undergo small, essentially random displacive movements as a result of their thermal energy. Such motions of individual particles may be likened to a kind of aimless three-dimensional *random walk*. If we limit our attention to the displacement of a single particle in a given direction from its position at some arbitrary zero of time, the *probability* that it will be found at a distance, $\pm x$, from its origin after a time, t , is given by:

$$P(x, t) = \frac{1}{2(\pi Dt)^{1/2}} \exp(-x^2/4Dt) \quad (1)$$

In this equation, the parameter D , called the *diffusion coefficient*, is a measure of the average rate with which the displacement of the particle occurs.

If, instead of attempting to follow the random motion of an individual particle, we introduce a large number of particles C_0 at a point within the system, their concentration $C(x, t)$ will vary with time and distance in a given direction according to:

$$C(x, t) = \frac{C_0}{2(\pi Dt)^{1/2}} \exp(-x^2/4Dt) \quad (2)$$

Both equations have the form of the well-known Gaussian error curve, which gives the distribution of random errors to be expected in a large set of measurements. Figure 1 illustrates these equations,

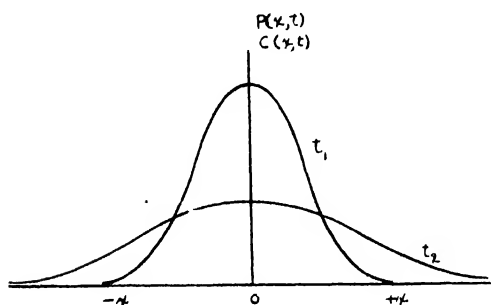


FIG. 1. Concentration or probability profiles for one-dimensional diffusion for times t_1 and t_2 .

showing the probability of finding a single particle at a distance from its origin (or the concentration distribution of a finite quantity of a substance) after times, t_1 and t_2 .

Equation (2) is a particular solution of a pair of more general differential equations known as Fick's laws. Consider a plane of unit cross-sectional area in a system, across which a concentration gradient, $\partial C/\partial x$, exists. There will be a net flux of matter, J_x , through the plane from the region of higher to lower concentration given by Fick's first law:

$$J_x = -D \frac{\partial C}{\partial x} \quad (3)$$

In principle, it is possible to determine the coefficient of diffusion on the basis of Eq. (3) by measuring the net quantity of matter that crosses through unit area of the plane in unit time. This proves to be difficult in liquids or solids, however, and an additional complication is that the concentration gradient $\partial C/\partial x$ is not constant but decreases with time.

For this reason, it is more feasible to measure the accumulation of matter in a small volume element after a measured period of time. We may consider such a volume element to be bounded by parallel planes of unit area, separated from one another by a distance dx , as shown in Fig. 2. The

rate of change of concentration within the volume element is expressed by Fick's second law:

$$\frac{1}{dx} \left[J_x - \left(J_x + \frac{\partial J}{\partial x} \cdot dx \right) \right] = \frac{dC}{dt} = D \frac{\partial^2 C}{\partial x^2} \quad (4)$$

This is the fundamental equation, upon which all experimental studies of diffusion depend.

One of the most widely used of the absolute methods for measuring diffusion coefficients of liquids is the "open capillary" technique, devised by Anderson and Saddington.¹ A capillary tube, usually less than 1 mm in diameter and 2-3 cm in length, is filled with an isotopically labeled substance and immersed in a thermostatted bath of the unlabeled liquid. Interdiffusion of the labeled and unlabeled molecules occurs across the open end of the capillary, and at the end of the experiment, the average concentration of labeled substance remaining in the capillary is determined by suitable radiochemical or mass spectrometric techniques. The relation between the diffusion coefficient and the initial and final concentration of labeled substance in the capillary is given by a series solution of Fick's second law:

$$\frac{C_{av}}{C_0} = \frac{8}{(\pi)^2} \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} \exp\left[-(2n+1)^2 \pi^2 Dt/L^2\right] \quad (5)$$

In this equation, for which L and t are, respectively, the capillary length and time, the series converges rapidly and as a rule may be terminated after the first and second term with negligible error.

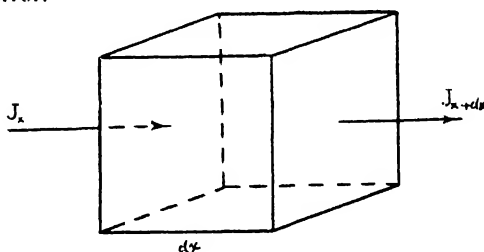


FIG. 2. Diffusive flux across parallel planes.

A completely different method for the measurement of diffusion coefficients in the liquid state is based upon nuclear magnetic resonance (NMR). An assembly of atomic nuclei which have been excited to some nonequilibrium spin distribution will return to thermal equilibrium by a mechanism that involves the coupling of the nuclear spins with their local molecular environment. The rate of the return to equilibrium is characterized by a *relaxation time* (T_1 , the spin-lattice relaxation time), and is governed by the variations in the local magnetic field at nuclei which are induced by the translational motions of neighboring molecules. Bloembergen, Purcell, and Pound² developed the basic theory which relates the spin-lattice

relaxation time to the diffusion coefficient of molecules, and made the first measurements on water and a series of hydrocarbons to test its validity.

Studies of the temperature dependence of the diffusion coefficient of liquids generally lead to an empirical relationship of the form:

$$D = D_0 \exp(-Q/RT) \quad (6)$$

where D_0 and Q are experimental parameters that are essentially temperature-independent and characterize the diffusion process in the system at hand. Because of the exponential form of Eq. (6), the mechanism of diffusion in liquids is often assumed to be an activated process, by analogy with other kinetic processes to which the *absolute reaction rate theory*³ has been applied with much success. According to this theory, the rate-limiting step of a kinetic process is determined by the frequency with which atoms or molecules acquire sufficient energy through thermal fluctuations to surmount an energy barrier identified by the parameter Q . It has not yet proved possible to make accurate *a priori* calculations of D_0 and Q on the basis of the activated state theory, however, and the exponential form of Eq. (6) is no proof that the mechanism of atom transport in liquids is a thermally activated process in any simple sense. At this time, it appears that a more fundamental insight into the structure and dynamic interactions of molecules in the liquid state will emerge from theories⁴ which concern themselves with the cooperative motions of particles that exchange momenta with one another during collisions. The success of such efforts will depend upon more accurate knowledge of intermolecular potentials.

The range over which diffusion coefficients of liquids⁵ vary is much more limited than that of solids (see DIFFUSION IN SOLIDS). As a rule, experimental values lie between 10^{-4} and 10^{-6} cm²/sec, but despite this relatively narrow compass, they contain information that must lead to a more complete description of the disordered liquid state.

NORMAN H. NACHTRIEB

References

1. Anderson, J. S., and Saddington, K., *J. Chem. Soc.*, 381 (1949).
2. Bloembergen, N., Purcell, E. M., and Pound, R. V., *Phys. Rev.*, 73, 679 (1948).
3. Glasstone, S., Laidler, K. J., and Eyring, H., "The Theory of Rate Processes," p. 477, New York, McGraw-Hill Book Co., 1941.
4. Rice, S. A., and Allnatt, A. R., *J. Chem. Phys.*, 34, 2144 (1961).
5. Jost, W., "Diffusion," p. 436, Academic Press, 1960.

Cross-references: BROWNIAN MOTION, DIFFUSION IN SOLIDS, KINETIC THEORY, MAGNETIC RESONANCE, RELAXATION.

DIFFUSION IN SOLIDS

The term "diffusion" refers to the random motion, generally activated by local fluctuations of thermal energy, of particles through a medium. The particles with which we shall be concerned are atoms and molecules; the medium can be various types of solids. Of special interest will be crystalline solids—these include all metals, most ionic substances, and many covalent ones—in which the atoms occupy periodic and well-defined sites.

The migrating particles may themselves be uniformly distributed constituents of the host solid; this is called self-diffusion. When the diffusing system contains chemical inhomogeneities or when a foreign substance diffuses in from the surface, we speak of chemical or of impurity diffusion. Diffusion processes are technologically important in the oxidation and tarnishing of metals, where one reactant must migrate through the layer of reaction product, and in the annealing of deformed or radiation-damaged materials. Self-diffusion is an essential step in the photographic process in silver halides, and impurity diffusion is widely used in the fabrication of semiconductor devices such as transistors. Many metals, such as steel and duralumin, are hardened by solid-state precipitation and reaction, in which diffusion plays a dominant role. It is also significant in the powder metallurgy technique of fabrication of parts from high-melting metals.

If, in the medium, there are variations in the concentration of the migrating atoms—perhaps chemically different atoms or, in the case of self-diffusion, radioactive tracer isotopes—then there occurs a net drift of the diffusing species from regions of high concentration to those of lower concentration. This flow takes place even though each individual atom may migrate completely at random. It is a statistical result of the fact that if there are more atoms per unit volume of, say, A to the left of a given plane than to the right, then even with random, nondirected motion, more A atoms will cross the plane from the left than from the right. We can define the flux of A as the net excess of A atoms crossing a plane of unit area in unit time. Experimentally, this flux is found to depend on the chemical natures of the medium and the diffusing species. If the medium is isotropic or is a crystal of cubic symmetry, the flux is along the direction of the concentration gradient. Moreover, if the system is not too thermodynamically nonideal, the flux is proportional to the concentration gradient; the constant of proportionality is called the diffusion coefficient, D . Thus, we write the flux $J = -D dc/dx$, where the negative sign indicates that the net flow is toward the region of lower concentration. This statement is known as Fick's law. If lengths are measured in centimeters and time in seconds, D is in units of square centimeters per second. It follows from Fick's law that at any given point in the medium, the concentration of the diffusing entity A will change with time at a rate governed by the variation of the flux with distance [$\partial c/\partial t = \partial/\partial x(D \partial c/\partial x)$]; for the particularly simple case where D is independent of

distance, as in self-diffusion, then $\partial c/\partial t \approx D \partial^2 c/\partial x^2$]. Also, from the theory of random flights, it can be shown that the root-mean-square displacement of atoms resulting from diffusion for a time t increases as the square root of the product Dt : $R_{rms} = (6Dt)^{1/2}$.

It is observed that for any given system, D increases rapidly with increasing temperature, almost invariably following the Arrhenius relation $D = D_0 \exp(-H/RT)$. Here, D_0 and H are positive constants for a given system and R is the universal gas constant. The parameter H is called the activation energy and generally increases as the melting point of the host crystal increases. Typically, in crystals which melt at 400 to 500°C, H is about 20,000 to 25,000 cal/mole, or 1 eV/atom, for self-diffusion or diffusion of substitutionally dissolved impurities. In crystals which melt near 1000°C, such as the noble metals, the activation energy is approximately 2 eV/atom. The value of the parameter D_0 is usually in the range 0.01 to 100 cm²/sec. It is interesting that for a large number of metals and simple ionic crystals, the diffusion coefficients for self-diffusion and for most impurities lie near to 10^{-8} cm²/sec at temperatures approaching the melting point. Thus, after diffusing for one day at such a temperature, the value of R_{rms} is about 1 mm, rather a large distance when compared to the spacing between atoms in a crystal.

Because of the three-dimensional regularity of atomic positions in a crystalline solid, the unit step in diffusion must be the jump of an atom from one site to a neighboring, crystallographically equivalent site. Large-scale diffusion is the result of random superposition of many such jumps, all of the same length λ but distributed among the various jump directions allowed by the crystal. It is readily shown that the relation between the macroscopic diffusion coefficient D and the microscopic atomic jump frequency Γ is $D = 1/6 \lambda^2 \Gamma$. This equation may be compared with that given above for R_{rms} by noting that for random jumps $R_{rms} = \lambda(\Gamma t)^{1/2}$. Now λ will depend on the details of the mechanism of diffusion but it must be of the order of the interatomic spacing, about 3×10^{-8} cm. Then the typical high-temperature diffusion coefficient of 10^{-8} cm²/sec requires each atom to make 10^7 to 10^8 jumps each second.

In most crystals the atoms are rather densely packed; thus the means whereby such a high frequency of jumps can be accomplished is not obvious. Conceptually, the simplest possibility is the simultaneous exchange of sites between two atoms, but this is ruled out because of the excessive activation energy that would be required to push aside the mutual neighbors of the pair. A dramatic demonstration that diffusion must proceed by a mechanism which allows independent motion of individual atoms is the Kirkendall effect. Two mutually soluble specimens of differing composition, say A and B, are welded together with inert markers imbedded at the interface. Subsequent interdiffusion of A and B results in a drift of the markers relative to the ends of the specimen,

indicating that more atoms have left one side of the couple than have entered it from the other. Clearly a pair exchange mechanism cannot be operative here.

Extensive evidence is now available that in most cases of self-diffusion or of substitutionally dissolved impurities, migration proceeds as a result of the presence and mobility of vacant lattice sites. These vacancies exist in the crystal in thermodynamic equilibrium, at concentrations which increase with temperature as $\exp(-H_f/RT)$, where H_f is the energy required to form a vacancy (about 1 eV in the noble metals). At temperatures near the melting point, the fraction of sites vacant is typically 0.01 to 0.1 per cent. Vacancies move by the jumping of adjacent atoms, at a rate which varies as $\exp(-H_m/RT)$. H_m is the activation energy for the migration process. The average jump frequency of an atom must then be the product of the jump frequency of a vacancy and the fraction of atomic sites that are vacant. Comparing the temperature dependence of the diffusion coefficient with that of these two factors, it follows that H must equal the sum of H_m and H_f . Quantitative experimental verification of this equality in a number of substances has firmly established the vacancy mechanism for diffusion in such crystals.

There are cases, however, where other mechanisms are known to operate. First, diffusing atoms migrate from one interstitial position (i.e., squeezed in between proper atom sites) to another when the migrating atom is very small (carbon in iron), when the host atoms and diffusing atoms are easily deformed (diffusion of silver ion in silver bromide), and occasionally in crystals which are not densely packed (copper in silicon). Second, when the temperature is so low that diffusion through the volume of the crystal is very slow, all substances show "short-circuiting" effects due to migration along external surfaces, grain boundaries, and line defects called dislocations. The details of these processes are incompletely understood at present.

LAWRENCE SLIFKIN

References

The first reference describes experimental techniques for determining diffusion coefficients; those following are recent brief reviews in order of increasing sophistication. References to more detailed discussions are given in these.

- Tomizuka, C. T., in Lark-Horovitz, K., and Johnson, V., Eds., "Methods of Experimental Physics," Vol. 6A, p. 364, New York, Academic Press, 1959.
- Girifalco, L. A., "Atomic Migration in Crystals," New York, Blaisdell, 1964.
- Shewmon, P. G., "Diffusion in Solids," New York, McGraw-Hill Book Co., 1963.
- Lidiard, A. B., "Ionic Conductivity," in "Handbuch der Physik," Vol. 20, p. 246, Berlin, Springer-Verlag, 1957.
- Lazarus, D., "Diffusion in Metals," in Seitz, F., and Turnbull, D., Eds., *Solid State Phys.*, 10, 71 (1960).

Cross-references: DIFFUSION IN LIQUIDS, SOLID-STATE PHYSICS, SOLID-STATE THEORY.

DIODE (SEMICONDUCTOR)

There exists a class of two terminal devices which have the property of permitting current to flow with practically no resistance in one direction and offer nearly infinite resistance to current flowing in the opposite direction. These devices are called *diodes*. The applications of diodes to electronic circuits are numerous. To mention a few, they include rectification of alternating current to a unidirectional current, detection of radio waves, and gating circuits used in digital computers.

The basic materials utilized for making semiconductor diodes are germanium (Ge) and silicon (Si). These elements are included in column IV of the periodic table (see PERIODIC LAW AND PERIODIC TABLE). Both Ge and Si are *tetravalent* elements, i.e., they have 4 valence electrons. Elements in their pure state are said to be *intrinsic*.

Each element under column III of the periodic table has 3 valence electrons and is referred to as *trivalent*. Examples of trivalent elements include indium (In) and gallium (Ga). Elements in column V of the table have 5 valence electrons and are called *pentavalent*. Arsenic (As) and antimony (Sb) are examples of pentavalent elements.

The process of introducing one of the elements from column III or V into intrinsic Ge or Si is called *doping*. The doped material becomes impure or *extrinsic*. If a trivalent impurity is introduced in Ge or Si (trivalent elements have one *less* valence electron than Ge or Si) holes are created and the material is said to be *p*-type. Introduction of a pentavalent impurity (pentavalent elements have one *more* valence electron than Ge or Si) creates free electrons and the material is *n*-type.

Because of thermal effects, free electrons and holes are always being produced in Ge and Si (intrinsic generation of electron-hole pairs). Consequently, there will be some electrons in the *p*-type material and some holes in the *n*-type material. These carriers are referred to as *minority* carriers. Electrons in *n*-type material and holes in *p*-type material are termed *majority* carriers.

There are two important kinds of semiconductor diodes: point-contact and *p-n* junction types. The operation of the point-contact diode is imperfectly understood whereas the operation of the junction diode is well known. The junction diode is most widely used; the point-contact type finds greatest use in high-frequency applications.

A typical point-contact structure is illustrated in Fig. 1. An S-shaped wire a few mils in diameter, called a cat's whisker, is pressed against a small wafer of doped semiconductor material. A point contact is formed. Assuming the semiconductor material is *n*-type Ge (or Si), a current of 100 to 200 mA is passed through the cat's whisker, point contact, and wafer for a length of time up to 100 msec. A *p-n* junction is formed and the point-contact diode obtained.

Considering the junction diode, imagine a

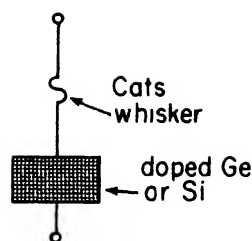


FIG. 1.

single crystal of Ge (or Si) doped so half the material is *p*-type and the other half, *n*-type. The internal boundary between the two extrinsic regions is a *p-n* junction, and the resulting device is a junction diode (Fig. 2). The electrical symbol for the point-contact and junction diodes is illustrated in Fig. 3.

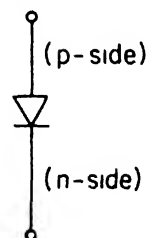
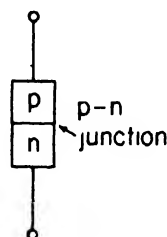


FIG. 3.

What are the characteristics of the *p-n* junction? To answer this question, three possible conditions are considered. Referring to Fig. 4(a), these are:

(1) *Unbiased*: *p*- and *n*-sides are connected by a wire.

(2) *Reverse biased*: the *p*-side is connected to the negative terminal of battery *E*, and the *n*-side connected to the positive terminal.

(3) *Forward biased*: the *p*-side goes to the positive terminal of *E*, and the *n*-side to the negative terminal.

Simple energy diagrams for the three conditions are shown in Fig. 4(b) for the electron. Similar diagrams can be generated for holes. When the diode is unbiased, no net flow of electrons takes place across the junction. Assuming that some electrons on the *n*-side have sufficient energy to overcome the potential hill, electrons on the

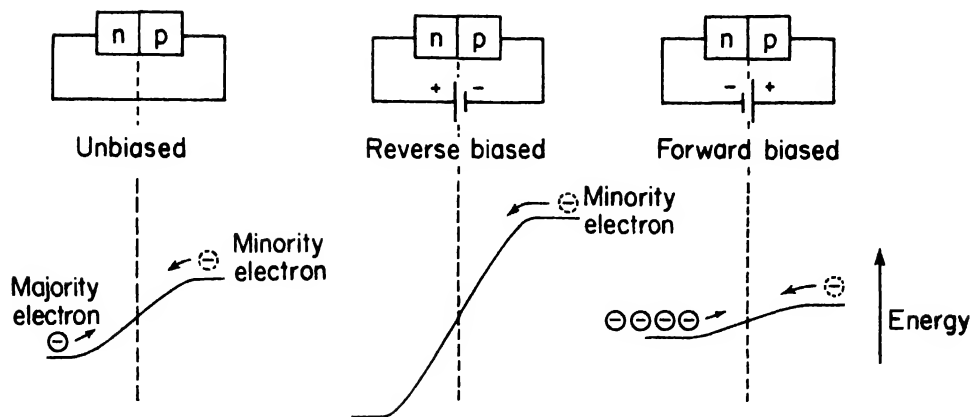


FIG. 4.

p-side (minority carriers) "slide down" the hill making the net current flow zero. For the reverse biased case, the potential hill is raised and only the few minority carriers from the *p*-side "slide down". This results in a minute reverse saturation current. When the diode is forward biased, the potential hill is lowered. This enables electrons to climb over the hill and current flow occurs. The same considerations apply to holes. In fact, the total diode current is equal to the sum of the electrons and holes flowing across the junction.

The characteristic curve of a semiconductor diode is shown in Fig. 5. An equation for this

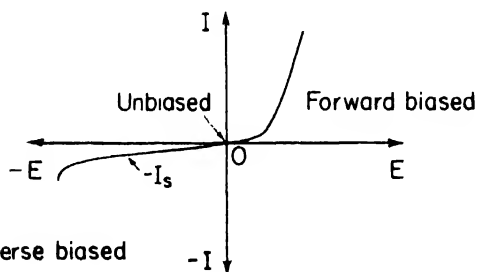


FIG. 5.

curve, called the *rectifier equation*, is expressed as:

$$I = I_s(e^{-11600 E / T} - 1)$$

where

I = diode current, amperes

I_s = reverse saturated current (which is temperature dependent), amperes

E = diode biasing voltage (+ E for forward bias; - E for reverse bias), volts

T = absolute temperature ($^{\circ}\text{C} + 273^{\circ}$), degrees Kelvin.

At room temperature (300°K) and $E > 0.1$ volt:

$$I \approx I_s e^{39 E}$$

When E is more negative than 0.1 volt:

$$I \approx -I_s$$

ARTHUR H. SEIDMAN

References

- Greiner, R. A., "Semiconductor Devices and Applications," New York, McGraw-Hill Book Company, Inc., 1961.
- Hunter, L. P., "Handbook of Semiconductor Electronics," second edition, New York, McGraw-Hill Book Company, Inc., 1962.
- Schwarz, R. F., "Introduction to Semiconductor Theory," *Electro-Technol.*, 107-130 (January, 1960).
- Seidman, A. H., "Solid-State Principles," *Electro-Technol* (Dec. 1964).
- Seidman, A. H., and Marshall, S. L., "Semiconductor Fundamentals: Devices and Circuits," New York, John Wiley & Sons Inc., 1963.
- Warschauer, D. M., "Semiconductors and Transistors," New York, McGraw-Hill Book Company, Inc., 1959.

Cross-reference: ENERGY LEVELS, POTENTIAL, TRANSISTOR.

DIPOLE MOMENTS (ELECTRICAL AND MAGNETIC)

Uncharged molecules can be classified as non-polar or polar depending on whether, in the absence of an electric field, the centers of gravity of their constituent positive and negative charges are coincident or not. A body containing two opposite charges, $\pm Q$, separated by a distance d , is characterized by an electric dipole moment, $Qd = \mu$; μ is a vector quantity, expressed conveniently in debye (D) units: 1 debye = 10^{-18} esu = 3.33×10^{-30} [coul m].

In the presence of an applied field a normally nonpolar molecule becomes dipolar by induction, i.e., by deformation of its electronic and atomic arrangements: $\mathbf{m} = (\alpha_e + \alpha_a)\mathbf{E}$, where the coefficients of proportionality α_e and α_a are

the electronic and atomic polarizabilities, respectively. In the general case of an anisotropically polarizable molecule, α_e and α_a are tensors, the components of which may be evaluated from observations of electric birefringence (Kerr effect), the depolarization of (Rayleigh) scattered light, refractive index dispersion, etc. An estimate of the mean of the three principal polarizabilities is given by $3R/4\pi N$, where R is a molecular refraction by the Lorenz-Lorentz formula; when R is extrapolated to infinite wavelength the mean polarizability obtained refers to the electronic deformations alone. Polarizabilities are expressed in volume units (cubic centimeters) (N = the avogadro number).

A field E exercises a torque on an electric dipole μ , tending to align it in the field direction in opposition to the randomness caused by thermal agitation. In a large assembly of molecules, therefore, a statistical and temperature-dependent equilibrium is achieved which corresponds to a slight excess of molecules having their permanent dipoles oriented antiparallel to the field so that the average moment \bar{m} of one molecule is apparently proportional to the field intensity, i.e., an orientation polarizability α_o is exhibited.

The electric dipole moment per unit volume of a dielectric material is the polarization vector P , understandable in magnitude as the charge density bound at the electrodes by a polarized dielectric. Based on the arguments of Mossotti (1850) and Clausius (1879), the polarization per mole is related to the dielectric constant ϵ by $M(\epsilon - 1)/d(\epsilon + 2) = 4\pi N\alpha/3$, where M/d is the molecular volume, N is the avogadro number, and α is the over-all polarizability. Debye (1912) showed α_o to be $\mu^2/3kT$ (k = Boltzmann's constant, T = absolute temperature) so that $\alpha = \alpha_e + \alpha_a + \alpha_o$, and the total polarization per mole τP is the sum of the electronic, atomic, and orientation polarizations: $\tau P = \epsilon P + \alpha P + \alpha_o P$. A possible fourth polarization mechanism, the blocking or trapping of migrating charge carriers in a dielectric, although ignored in the classical molecular theory, may also contribute to the apparent ϵ of solid or macromolecule-containing systems.

The commonest method for the determination of dipole moments involves the dispersion of ϵ : τP is measured at radio and optical wavelengths (the second of these is a molecular refraction since the square of the index of refraction of a nonabsorbing, nonmagnetic material equals the dielectric constant at the same frequency), then approximately $R = \epsilon P + \alpha P$, and $\mu^2 = 9kT(\tau P - R)/4\pi N$. Although strictly valid only for gaseous dielectrics the Mossotti-Clausius-Debye equations have proved applicable also to solutes in nonpolar solvents, and by using alligation formulas, values of τP for a dissolved species can be obtained at infinite dilution; such estimates are usually close to, but not identical with, the true τP 's directly observed on the vaporized solutes. Over the past thirty years, much effort has been devoted to theoretical or empirical

treatments of "solvent effects." Since distortion polarizations are almost invariant with temperature, the temperature dependence of τP follows as $(\tau P)_T = A + B/T$; the constants A and B , when fitted to experimental data by least squares, give $A = \epsilon P + \alpha P$ and $B = 4\pi N\mu^2/9k$, whence $\mu = 0.012812B^{0.5}$ esu; results for about 350 gases are listed by Marryott and Buckley.

Practical details concerned with the measurement of dielectric constants, and other properties, necessary for the deduction of μ 's of solutes or vapors, are described fully in the books (cited below) by Le Fèvre, Smith, and Smyth, wherein also references are made to other, but less simple, techniques by which dipole moments can be determined (e.g., Stark splitting in microwave spectra of gases at low pressures, the dielectric losses or power factors of dilute solutions, molecular beam studies, etc.); the first two of these are useful since they can detect very small moments which the ordinary dielectric constant methods cannot reveal accurately; the third technique—involving the deviation undergone by a thin ribbon of gaseous molecules in passing through an intense nonhomogeneous electric field—is applicable to substances, such as metal salts, which through insolubility or low volatility would be otherwise unexaminable.

By the end of 1961, some 7000 dipole moment values for more than 6000 substances had been recorded (see McClellan's Tables); they fall mostly in the range 0 to 5 debyes.

Chemical interest is largely due to the relationships between polarity and molecular structure. Monatomic molecules, diatomic molecules of the type AA , and centrosymmetric polyatomic molecules, are nonpolar; a linear triatomic molecule ABA is nonpolar, but if bent or constructed as AAB it is polar; pyramidal tetraatomic molecules AB_3 are polar, etc. A more quantitative approach supposes that characteristic polarities are associated with covalent chemical bonds, e.g., that two bonds, having "bond moments" μ_1 and μ_2 , mutually inclined at θ° , produce a resultant of $(\mu_1^2 + \mu_2^2 + 2\mu_1\mu_2 \cos \theta)^{0.5}$. On this basis, bond moments deduced from the resultant moments of molecules with known structures, often permit the discovery or testing of stereo specifications of further molecules. However, caution is necessary since bond moments are not independent of bond environments, but may be modified by induced moments—determined by the fields of neighboring polar bonds or centers and the (anisotropic) polarizabilities of the bonds under consideration—or by other internal electronic effects (c.f. resonance, mesomerism, hybridization, etc.). Completely successful calculations of dipole moments from *a priori* theory have yet to be made.

Some concepts developed for electrostatic fields have magnetic counterparts; thus in place of polarization P there is magnetization I , the magnetic dipole moment per unit volume caused in a material by an externally applied field H ; internally the magnetic flux density (the magnetic induction) is $B = H + 4\pi I$; the ratio I/H is the

volume susceptibility κ ordinarily measured. Individual magnetic monopoles are not known to exist in nature, but movements and spins of electrons in atoms and molecules—if viewed classically as direct currents flowing in closed circuits—can create fields identical with those expected from magnetic dipoles having moments dimensionally equivalent to products of pole strengths and distances. The elementary magnetic moment is the "Bohr magneton," 9.273×10^{-21} [erg gauss⁻¹], assumed to be the magnetic moment of an electron "spinning" on its own axis. Atoms may possess orbital moments (due to mechanical angular movements of electrons) and spin moments (one for each electron). Magnetic moments can be induced or permanent. A unit volume containing ν particles each of magnetizability α_m , subjected to a field H , displays a magnetization $I = \nu\alpha_m H = \nu\bar{m}$, where \bar{m} is the average magnetic dipole moment per particle; thus $\kappa = \nu\alpha_m$, and the molar susceptibility $\chi = \kappa V = N\alpha_m$ (where V is the molar volume and N the avogadro number); α_m can be split into $\alpha_i + \alpha_p$, to correspond with the contributions to I made by the induced and permanent moments respectively.

An electron, in an orbit of radius r represents a current loop; application of a magnetic field H perpendicularly to the loop plane will induce a voltage tending to create a field opposing that applied; the effect will be manifest as an apparent induced moment antiparallel to H and—by classical calculations—of the value $-e^2 r^2 H / 6mc^2$ (here e is the electronic charge, m is the electronic mass, c is the velocity of light); hence $\alpha_i = -e^2 r^2 / 6mc^2$, and for a monatomic substance with spherical atoms the molar diamagnetic susceptibility $\chi = -(Ne^2 / 6mc^2) \sum_n r_i^2$, where r_i^2 is the mean value of r^2 for the i th electron and the sum is taken over n electrons. The χ 's observed for the inert gases, the C atoms in diamond, the Cl atoms in Cl_2 , etc. have agreed with reasonable magnitudes of $\sum r^2$. Pascal (1910) showed diamagnetic susceptibility to be an "additive-constitutive" property, so that the χ 's of polyatomic molecules can be approximately predicted by summing "atom" and "bond" susceptibilities in numbers and kinds appropriate to the molecular structure under consideration. The diamagnetic susceptibility of an individual molecule is a tensor quantity; χ/N by experiment is an average of three principal magnetic susceptibilities directed along three mutually perpendicular principal axes of magnetic susceptibility; these can be investigated through torsional movements of crystals in magnetic fields (Krishnan's method) or from magnetic birefringence measurements (Cotton-Mouton effects) in conjunction with data for χ_{mean} secured with a Gouy balance.

A permanent magnetic dipole will experience a torque in a magnetic field. Langevin (1905) showed that the mean moment \bar{m} of a gaseous molecule in the field direction (provided that H is not too large) is $\bar{m} = (m_p^2 / 3kT)H$, where m_p is the actual moment of each molecule; therefore the molar paramagnetic susceptibility χ_p is

$N\bar{m}/H = N\alpha_p = Nm_p^2 / 3kT$. In practice χ_p is extracted from the observed χ by treating this as the algebraic sum of a negative diamagnetic susceptibility (estimated from Pascal's constants) and a positive paramagnetic susceptibility; thus m_p follows as $(3kT\chi_p/N)^{0.5}$ [erg gauss⁻¹] or as $2.84(T\chi_p)^{0.5}$ [Bohr magnetons]. Molar diamagnetic susceptibilities are independent of temperature, while molar paramagnetic susceptibilities in general vary as $1/T$ or $1/(T - T_c)$. The small paramagnetisms of alkali metals, Cu, Ag, etc., or of certain salts (e.g., KMnO_4 or $\text{K}_2\text{Cr}_2\text{O}_7$), attributable respectively to uncompensated spins of conduction electrons, or to uncompensated paramagnetisms of complex ions, are temperature invariant.

Normally any atom or molecule with unpaired electrons shows paramagnetism and possesses a magnetic moment. Magnetic properties can therefore provide important information on valency states in free radicals, molecules containing first period elements with unpaired p electrons, transition elements having unpaired d electrons, lanthanides with unpaired $4f$ and actinides with unpaired $5f$ electrons. Theoretical expressions exist to calculate paramagnetic moments in terms of atomic structures and spin and orbital angular momenta of unpaired electrons. Simple examples are the ions of transition metals where, if n is the number of unpaired electrons, m_p is approximately predicted as $[n(n + 2)]^{0.5}$ Bohr magnetons (for a full discussion of such relations see Nyholm's review). Determination of m_p thus gives n , which is often of value in deciding the three-dimensional arrangements and bond types involved in molecules, especially those built around a central metal atom.

For most substances χ is independent of field strength, but a few paramagnetic compounds can, below the characteristic temperature T_c (see above), show "ferromagnetism" due to spontaneous parallel alignments of spins of atomic magnets. Materials which are ferromagnetic at ordinary temperatures (e.g., soft iron) have nonlinear magnetization-field characteristics, develop large magnetizations in weak fields, rapidly approach saturation conditions, exhibit hysteresis, etc.; whole domains about 0.01 mm in diameter and magnetically saturated are thought to be undergoing orientation during such processes. Ferromagnetism—and related phenomena such as "antiferromagnetism" and "ferrimagnetism"—have at present few applications in chemistry; in electronics (e.g., ferrites in antennas and in magnetic tape), they are frequently important.

R. J. W. LE FÈVRE

References

- Debye, P., "Polar Molecules," New York, The Chemical Catalog Co., Inc., 1929.
- Hippel, A. R. von, "Dielectrics and Waves," New York, J. Wiley & Sons, Inc., 1954.
- Le Fèvre, R. J. W., "Dipole Moments," Third edition, London, Methuen and Co., Ltd., 1953.

McClellan, A. L., "Tables of Experimental Dipole Moments," San Francisco and London, Freeman and Co., 1963.

Maryott, A. A., and Buckley, F., "Table of Dielectric Constants and Electric Dipole Moments of Substances in the Gaseous State," *Natl. Bur. Std. Circ.*, 537 (1953).

Nyholm, R. S., *Quart. Rev. London*, 7, 377 (1953).

Selwood, P. W., "Magnetochemistry," New York, Interscience Publishing, 1956.

Smith, J. W., "Electric Dipole Moments," London, Butterworth's Scientific Publications, 1955.

Smyth, C. P., "Dielectric Behavior and Structure," New York, Toronto, London, McGraw-Hill Book Co., Inc., 1955.

Van Vleck, J. H., "The Theory of Electric and Magnetic Susceptibilities," Oxford, Clarendon Press, 1932.

Cross-references: BOND, CHEMICAL; DIELECTRIC THEORY; FERROMAGNETISM; MAGNETISM.

DOPPLER EFFECT

The wave effect by which astronomers measure the radial velocities of galaxies, and policemen determine the speeds of approaching automobiles, was in spite of its simplicity not discovered until the nineteenth century. In 1842, Christian Doppler predicted that the frequencies of received waves were dependent on the motion of the source or observer *relative to the propagating medium*. His predictions were promptly checked for sound waves by placing the source or observer on one of the newly developed railroad trains.

In his original article on the special theory of relativity (see RELATIVITY), Einstein¹ developed the expression for the Doppler shift of light waves which was dependent upon the velocity of the source *relative to the observer*. From the photon hypothesis for light, Schrödinger^{2,3} obtained the same results. Thus, the Doppler effect provides one of the illustrations of the equivalence of the wave and particle descriptions of light.

Classroom demonstrations of the Doppler effect for water waves are made in shallow glass-bottom ripple tanks. Instead of giving the vibrating source a constant velocity, one lets the sheet of water as medium flow continuously by the source.

The circles of Fig. 1 are snapshots of the crests of a water wave or compressions in a sound wave observed when the source is moving at constant velocity v to the right relative to the medium. Points 1 and 2 are positions of the source one and two periods after passing O. The largest circular crest originated at O, the next at 1 and the smallest at 2. A crest is about to leave point 3 at the time the snapshot is taken. If the position P of the observer is a large distance from the source compared to the distance the source moves in one period, then with good approximation we may assume that two successive crests are moving in the same direction as they pass P. If v is the velocity of the source in the direction OA, the source moves a distance vT in one period T . In

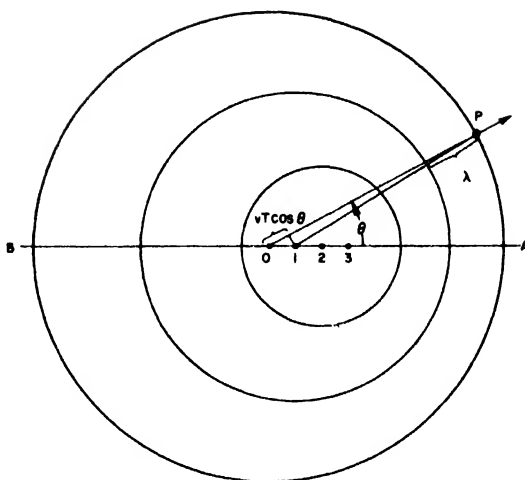


FIG. 1. The source is moving relative to the medium along the line BA. The circles represent crest of the wave at an instant. (Andrews, C. L., "Optics of the Electro-Magnetic Spectrum", Englewood Cliffs, N.J., Prentice-Hall, Inc., 1960).

one period, the source comes closer to P by the amount $vT \cos \theta$, where θ is the angle between the direction of the velocity of the source and the line from the source to the observer at P. Now λ_0 is the wavelength and ν_0 the frequency when the source is at rest; λ is the observed wavelength and ν the observed frequency when the source is in motion. Because of the motion of the source, the wavelength received at P is reduced by $vT \cos \theta$.

$$\lambda - \lambda_0 = vT \cos \theta$$

If c is the velocity of the wave, $T = \lambda_0/c$ and $\lambda = \lambda_0 \left(1 - \frac{v}{c} \cos \theta \right)$, but $\lambda = c/\nu$ and $\lambda_0 = c/\nu_0$. Therefore,

$$\frac{\nu}{\nu_0} = \frac{1}{1 - \frac{v}{c} \cos \theta} \quad (1)$$

when the source is in motion relative to the medium.

In Fig. 2 the source is at rest, but the observer at P has a velocity v with respect to the medium. The velocity of the wave relative to the observer is equal to the vector sum of the velocity of the wave relative to the medium and the velocity of the medium relative to the observer. In Fig. 2, c is the velocity of the wave and v the velocity of the observer relative to the medium. Let λ_0 be the wavelength, ν_0 the frequency of the source, and ν the frequency received by the moving observer. The radial velocity of the wave relative to the observer is $c + v \cos \theta$ so that

$$\nu \lambda_0 = c + v \cos \theta$$

For an observer at rest $\nu_0 = c/\lambda_0$. Substituting for λ_0 , we obtain

$$\frac{\nu}{\nu_0} = 1 + \frac{v}{c} \cos \theta \quad (2)$$

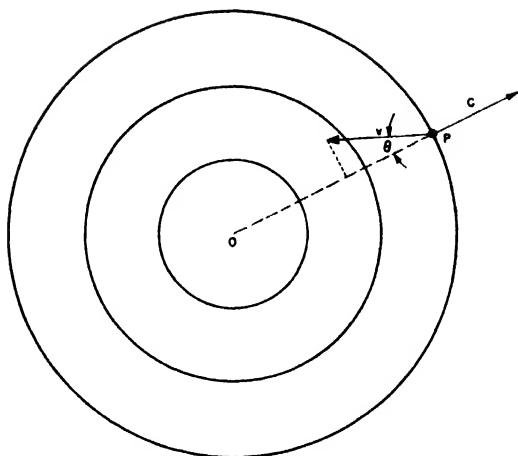


FIG. 2. The source is at rest at point O and the observer at point P is moving with velocity v relative to the medium. (Andrews, C. L., "Optics of the Electro-Magnetic Spectrum," Englewood Cliffs, N.J., Prentice-Hall, Inc., 1960).

when the *observer* is in motion relative to the medium.

By a postulate of relativity, the velocity of light is the same relative to all observers. The theory of relativity yields the frequency

$$\frac{\nu}{\nu_0} = \frac{1 + \frac{v}{c} \cos \theta_0}{\sqrt{1 - \frac{v^2}{c^2}}} \quad (3)$$

in which $v \cos \theta_0$ is the component of the velocity of the source toward the observer. The angle θ_0 is measured in the source system. If θ is the angle measured in the observers system, then

$$\cos \theta_0 = \frac{\frac{v}{c} \cos \theta}{\frac{v}{c} \cos \theta - 1} \quad (4)$$

Figure 3 is a graphical plot of ν/ν_0 against v/c for the radial motion in the three cases we have treated. (1) The linear relation is that for the observer in motion relative to the medium that propagates sound or other mechanical waves. (2) The other solid curve is for the source of sound in motion. (3) The broken curve represents the Doppler effect for electromagnetic waves such as x-rays, light, and radio waves.

By comparing several spectral lines of elements observed in a star with a laboratory spectrum of the same elements, astronomers use the Doppler effect to measure the radial components of velocity of astronomical bodies toward or away from the earth. Spectra of the edges of the sun's disk are measured to determine the velocities toward and away from the earth. The radial velocities of the principal stars of our galaxy have

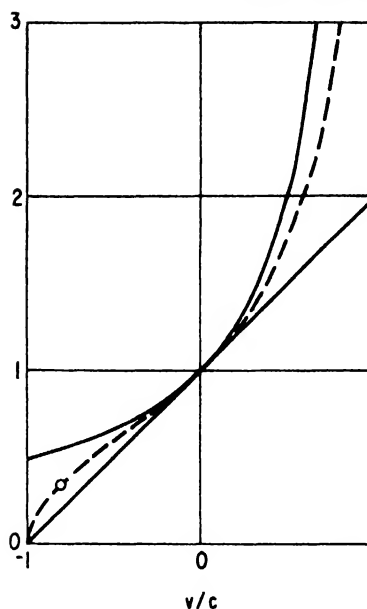


FIG. 3. Graphical plots of the ratio of the observed frequency to the frequency at the source against the ratio of radial velocity to the velocity of the wave for three cases: (1) sound waves from a moving source, (2) sound waves to a moving receiver, (3) electromagnetic waves. The circle represents the red shift of light received from the most distant galaxies observed.

been recorded. The spectral lines of some of the stars are doublets which periodically come together and separate again indicating that the light comes from two stars revolving about a common center of gravity (see ASTROMETRY).

In the expanding universe, the radial velocities of other galaxies away from our galaxy are proportional to their distances from the observer. Thus, the Doppler red shift provides a means of determining the dimensions of the observed universe. In 1964, some of the most intense radio sources (see RADIO ASTRONOMY) were located with high precision by observing these sources when the moon passed in front of them. With this knowledge of position, the same sources were located with a light telescope.⁴ The measured red shift was surprisingly high. The sources were not stars as previously thought but the most distant galaxies known. One of them, 3C-9 in the catalogue of radio sources had a red shift $\Delta\lambda/\lambda_0$ equal to 2.0. If this shift were due solely to the Doppler effect, the astronomers had to conclude that the source was moving from us with 80 per cent of the velocity of light. The Doppler frequency and velocity of this source is indicated by a circle on the broken line of Fig. 3.

If a microwave beam is reflected from a moving microwave mirror, such as a person, an automobile or a man-made satellite, the image of the primary source may be considered as another source moving with twice the velocity of the mirror. Since the speed is small compared with the

speed of light, the squared terms of Eq. (3) may be neglected. Thus

$$\nu = \nu_0 \left(1 + \frac{2v}{c} \cos \theta \right)$$

Direct frequency measurements cannot be made to enough significant figures to distinguish ν from ν_0 . However, if the two frequencies are combined they give beats or the difference frequency

$$\Delta\nu = \nu_0 \frac{2v}{c} \cos \theta$$

Since the beat frequency is proportional to the radial velocity, a frequency meter may be calibrated in miles per hour. The precision of such a speed detector depends upon the frequency of the source being so stable that it varies less than the Doppler frequency shift during the time that the wave travels from the source to the mirror and back. The same phenomena of beats between the direct wave from the source and the wave reflected from a moving mirror may be observed with light. If one of the mirrors of Michelson's interferometer is moved at constant speed, the frequency with which dark bands pass the cross hair is the difference in frequency of the two waves.

The numerator of Eq. (3) contains a term for the radial component of velocity. However, the second-order term in the denominator is independent of direction. Thus, as v/c approaches unity, one may expect to detect a *tangential Doppler effect*. Ives and Stilwell⁵ have measured the predicted value for the Doppler shift in frequency due to a stream of radiating molecules for which v/c was 10^{-2} . This experiment was a direct proof of time dilatation (see RELATIVITY) for the transverse case. In order to separate the tangential from the radial effect, Ives and Stilwell produced a sharply collimated beam of molecules. In order that θ be precisely 90° , they set a mirror accurately normal to the line of observation and altered the line of observation until nearly the same wavelengths were given by direct and reflected light.

C. L. ANDREWS

References

1. Einstein, A., *Ann. Physik*, **17**, 891 (1905).
2. Schrödinger, E., *Physik. Z.*, **23**, 301 (1922).
3. Michels, W. C., *Am. J. Phys.*, **15**, 449 (1947).
4. Greenstein, J. L., and Schmidt, M., *Astrophys. J.*, **140**, 1 (1964).
5. Ives, H. E., and Stilwell, A. R., *J. Opt. Soc. Am.*, **31**, 369 (1941).

Cross-references: ASTROMETRY, RADIO ASTRONOMY, RELATIVITY, WAVE MOTION.

DYNAMICS

Introduction. A recent Webster's definition of the word *dynamics* is "That branch of MECHANICS treating of the motion of bodies (kinematics) and

the action of forces in producing or changing their motion (kinetics)."^{*}

This order of presentation is often utilized in textbooks and other treatises on the subject. With various interpretations of the word *bodies*, the subject has also been divided into the areas of particle dynamics, rigid body dynamics, and fluid dynamics. As needs have developed, various specializations have been created and new theories have been formulated. These special areas include mechanical vibrations, flight dynamics, space dynamics, gas dynamics, magnetohydrodynamics, and relativistic dynamics, to mention a few.

The present state of knowledge recognizes all motion as relative, since no fixed reference is known for finding the absolute motion of any body. The earth, which is frequently used as a frame of reference is rotating about its own axis and is also revolving about the sun. The solar system, which consists of the sun and its planets, is a minute part of the Milky Way galaxy that is known to be revolving in space.¹ Beyond this, there is limited knowledge of the nature of motion that exists.

In the field of astronomy, measurements are evaluated in a coordinate system that is located relative to the *fixed* stars. These stars are located at such a vast distance from the earth that they appear as points of light that are almost motionless in space. In this frame of reference, the motions of celestial bodies are described with extremely great precision, and the motions of bodies within the solar system can be predicted accurately over periods of hundreds of years.

Some applied areas of dynamics, exemplified by the space exploration program[†], also require the degree of extreme accuracy that is possible with a celestial frame of reference. In many other areas, this extreme accuracy is not essential and measurements based upon this frame of reference would be tedious and impractical. In such cases, motion may often be adequately described in a coordinate system located relative to the earth.

Kinematics. A particle is a body having dimensions that are small, relative to other dimensions of the system, so that its motion may be considered equivalent to the motion of a point at its mass center with rotational effects neglected. Thus, particles may be either small or large. In the solar system, it would be possible to assume the earth to be a particle but, in a terrestrial system, this assumption could be totally unjustifiable.

A rigid body is a group of particles having unvarying external and internal configuration. The size of the rigid body would be appreciable in comparison to the other dimensions of the system, so that the rotational effect would have to be considered.

A fluid body is a group of particles with varying external and or internal configuration.

^{*} By permission. From Webster's New Collegiate Dictionary, copyright 1956 by G. & C. Merriam Co., Publishers of the Merriam-Webster Dictionaries.

The analysis of this type of system will not be considered in this article. (See FLUID DYNAMICS.)

The kinematic analysis of the motion of a particle may be approached through the establishment of a position VECTOR \mathbf{r} , directed from the origin of a specified fixed coordinate system to the point representing the position of the particle, to give

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k} \quad (1)$$

where \mathbf{i} , \mathbf{j} and \mathbf{k} are unit vectors along the x , y and z axes, respectively. Differentiating Eq. (1) with respect to time yields a velocity equation of the form

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k} \quad (2)$$

A second differentiation with respect to time yields the acceleration of the particle in the form

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d^2\mathbf{r}}{dt^2} = \frac{d^2x}{dt^2}\mathbf{i} + \frac{d^2y}{dt^2}\mathbf{j} + \frac{d^2z}{dt^2}\mathbf{k} \quad (3)$$

If x , y and z are scalar functions of time, then Eq. (1), (2) and (3) may be used to trace the path of the particle and determine the velocity and acceleration at any instant. These equations may be easily adapted to the cases of rectilinear translation and curvilinear translation of a particle in plane motion.

The kinematic analysis of a rigid body moving in a plane often involves the trace of two points, which may be called A and B, that are located on the body. These points move with the body and remain a fixed distance apart. Two coordinate systems may be used to define the position of the body in the plane. An X - Y coordinate system, with origin O , is a fixed reference, and an x - y coordinate system, with origin o located at A, is attached to the body so that it moves and rotates with the body. In the fixed reference system, a position vector \mathbf{R} is directed from O to point A on the body and a second position vector $\boldsymbol{\rho}$ is directed from O to point B. In the moving coordinate system, a vector \mathbf{r} is directed from A to B. An equation relating the position of the two points may be written as

$$\boldsymbol{\rho} = \mathbf{R} + \mathbf{r} \quad (4)$$

Using the \mathbf{I} , \mathbf{J} , \mathbf{K} unit vectors for the X - Y - Z coordinate system and the \mathbf{i} , \mathbf{j} , \mathbf{k} unit vectors for the x - y - z coordinate system, Eq. (4) may be rewritten as

$$\boldsymbol{\rho} = X\mathbf{I} + Y\mathbf{J} + x\mathbf{i} + y\mathbf{j} \quad (5)$$

Differentiating Eq. (5) with respect to time yields the velocity of B in the form

$$\mathbf{v}_B = \frac{d\boldsymbol{\rho}}{dt} = \frac{dX}{dt}\mathbf{I} + \frac{dY}{dt}\mathbf{J} + \frac{dx}{dt}\mathbf{i} + x\frac{d\mathbf{i}}{dt} + \frac{dy}{dt}\mathbf{j} + y\frac{d\mathbf{j}}{dt} \quad (6)$$

which simplifies to

$$\mathbf{v}_B = \mathbf{v}_A + \mathbf{v}_{B/A} \quad (7)$$

where \mathbf{v}_B and \mathbf{v}_A are the velocities of points B and A, respectively, and $\mathbf{v}_{B/A}$ is the velocity of point B relative to point A.

A second differentiation of Eq. (5) with respect to time yields the acceleration of B in the form

$$\mathbf{a}_B = \mathbf{a}_A + \mathbf{a}_{B/A} \quad (8)$$

where \mathbf{a}_B and \mathbf{a}_A are the accelerations of points B and A, respectively, and $\mathbf{a}_{B/A}$ is the acceleration of point B relative to point A.

A second important case of rigid body motion exists when point B is not attached to the same body as point A but is moving along a constrained path on this body. For the analysis of this motion, it is convenient to designate the fixed reference as body 1, the body to which point A is attached as body 2, and the body to which point B is attached as body 3. The same general arrangements of coordinate systems are used but in this case, the fixed X - Y coordinate system may be considered as attached to body 1 while the moving x - y coordinate system is attached to body 2. In the general case, vector \mathbf{r} within the x - y coordinate system is varying in both magnitude and direction. Differentiating the position expression

$$\boldsymbol{\rho} = \mathbf{R} + \mathbf{r} \quad (9)$$

once with respect to time and simplifying yields

$$\mathbf{v}_{B3} = \mathbf{v}_{A2} + \mathbf{v}_{B2/A2} + \mathbf{v}_{B3/2} \quad (10)$$

where \mathbf{v}_{B3} is the velocity of point B on body 3, \mathbf{v}_{A2} is the velocity of point A on body 2, $\mathbf{v}_{B2/A2}$ is the velocity of point B on body 2 relative to point A on body 2, and $\mathbf{v}_{B3/2}$ is the velocity of point B on body 3 relative to body 2.

A second differentiation with respect to time gives

$$\mathbf{a}_{B3} = \mathbf{a}_{A2} + \mathbf{a}_{B2/A2} + \mathbf{a}_{B3/2} + 2\boldsymbol{\omega}_2 \times \mathbf{v}_{B3/2} \quad (11)$$

where \mathbf{a}_{B3} is the acceleration of point B on body 3, \mathbf{a}_{A2} is the acceleration of point A on body 2, $\mathbf{a}_{B2/A2}$ is the acceleration of point B on body 2 relative to point A on body 2, $\mathbf{a}_{B3/2}$ is the acceleration of point B on body 3 relative to body 2, $\boldsymbol{\omega}_2$ is the angular velocity of body 2, and $\mathbf{v}_{B3/2}$ is the velocity of point B on body 3 relative to body 2. The term $2\boldsymbol{\omega}_2 \times \mathbf{v}_{B3/2}$ is often referred to as the CORIOLIS component of acceleration.

Equations 7 and 8 may be adapted to the case of rotation of a rigid body about a fixed axis at point A by considering point A to be fixed. Thus, \mathbf{v}_A and \mathbf{a}_A are both zero and

$$\mathbf{v}_B = \mathbf{v}_{B/A} \quad (12)$$

$$\mathbf{a}_B = \mathbf{a}_{B/A} \quad (13)$$

For a body rotating about a fixed axis, analysis of the rotational motion yields

$$\frac{d\theta}{dt} = \omega \quad (14)$$

$$\frac{d\omega}{dt} = \alpha \quad (15)$$

where θ is the angular displacement in radians, ω is the angular velocity in radians per second, and α is the angular acceleration in radians per second per second. It should be noted that time may be expressed in other units.

Kinetics—Newton's Laws of Motion. The laws of Newton are based upon the motion of a particle relative to a fixed frame of reference in which the particle can be made completely free of all outside influences. Under such a condition, the particle at rest will remain at rest and a particle in motion will continue to move at a constant velocity. This ideal frame of reference is often referred to as a Newtonian or as an inertial frame of reference.²

Since it is not possible to actually establish the Newtonian frame of reference, Newton's laws of motion are used in a celestial or a terrestrial frame of reference. The gyroscopic instruments and the stabilized platforms represent attempts to achieve a fixed or stabilized frame of reference for aircraft or space vehicles.

In modern terminology, Newton's laws of motion for a particle may be interpreted as

(1) A particle tends to remain at rest or continues to move at a constant velocity if there is no unbalanced force acting upon it.

(2) An unbalanced force acting on a particle will produce a time rate of change of momentum which, at any instant, will be proportional to the force and will be in the same direction as the force.

(3) The forces that exist between two contacting particles are equal in magnitude, are opposite in direction, and are collinear.³

The concept of the first law is the fundamental principle used for the analysis of forces acting on stationary particles and also for the analysis of forces acting on particles moving with a constant velocity. The concept has been expanded to include rigid bodies and fluid bodies.

The second law is the foundation of the analysis of forces acting on particles moving with accelerations and, again, it has been extended to include rigid bodies which involve rotary motion and to include fluid bodies.

The third law is fundamental to the force analysis of interconnecting systems of particles under both static and dynamic conditions. It has also been extended to include simple contact between any pair of bodies and, with some modification, to include any type of interaction between bodies.

Force and Acceleration. In general, Newton's second law is stated as

$$\Sigma \mathbf{F} = k \frac{d(m\mathbf{v})}{dt} \quad (16)$$

where $\Sigma \mathbf{F}$ is the net unbalanced force on the particle; k is a constant of proportionality, consistent with the units used, that is determined

experimentally; m is the mass of the particle; and \mathbf{v} is the instantaneous velocity of the particle.

For a particle that is not shedding or accumulating mass, Newton's second law reduces to

$$\Sigma \mathbf{F} = k m \mathbf{a} \quad (17)$$

where $\Sigma \mathbf{F}$, k , and m are as previously defined and \mathbf{a} is the instantaneous acceleration of the particle. By the proper choice of units in Eq. (16) and (17), the constants of proportionality can be made equal to unity.

For a rigid body in plane motion, both translational and rotational acceleration must be considered. By extending the concept of Newton's second law, it may be stated that

$$\Sigma \mathbf{F} = k_1 m \mathbf{a} \quad (18)$$

$$\Sigma \mathbf{T} = k_2 I \alpha \quad (19)$$

where $\Sigma \mathbf{F}$ and $\Sigma \mathbf{T}$ are the unbalanced force and unbalanced torque, respectively, acting on the body; k_1 and k_2 are constants of proportionality, consistent with the units used, that are determined experimentally; m is the mass of the body; I is the mass moment of inertia of the body about a centroidal axis that is perpendicular to the plane of the motion; \mathbf{a} is the linear acceleration of the center of mass of the body; and α is the angular acceleration of the body. Again, with proper choice of units, k_1 and k_2 can be made equal to unity.

The concepts and equations presented herein can be applied for the solution of a wide variety of problems which involve systems of particles, systems of bodies, or combinations thereof. The usefulness of this approach may be further broadened by the introduction of the closely related concepts of energy and impulse—momentum.

It should be noted that, in Newtonian mechanics, the fundamental property of the particle is an unvarying mass and time is absolute. It should also be noted that when the speed of the particle approaches the speed of light, this theory becomes inaccurate in comparison to a theory based upon a more exact mathematical model attained through the application of the principles of RELATIVITY.⁴

GLENN L. DOWNEY

References

1. Robertson, H. P., "The Universe," *Sci. Am.*, 195, No. 3, 73-81 (September, 1956).
2. Goodman, L. E., and Warner, W. H., "Dynamics," Belmont, Calif., Wadsworth, 1964.
3. Downey, G. L., and Smith, G. M., "Advanced Dynamics," Scranton, Pa., International Textbook, 1960.
4. Synge, J. L., and Griffith, B. A., "Principles of Mechanics," Third edition, New York, McGraw-Hill, 1959.

E

ELASTICITY

Elasticity is the part of mechanics dealing with deformations that vanish entirely once the forces that have caused them are removed. Most solid bodies behave elastically for sufficiently small deformations, and we will be concerned here with the infinitesimal theory of elasticity. Also we will consider only isotropic bodies, that is, bodies whose elastic properties are the same in all directions.

The fundamental quantities in elasticity are second-order tensors, or dyadics: the deformation is represented by the *strain dyadic*, and the internal forces are represented by the *stress dyadic*. The physical constitution of the deformable body determines the relation between the strain dyadic and the stress dyadic, which relation is, in the infinitesimal theory, assumed to be linear and homogeneous. While for anisotropic bodies this relation may involve as much as 21 independent constants, in the case of isotropic bodies, the number of elastic constants is reduced to two.

Let $\mathbf{s}(\mathbf{r})$ be the displacement vector, due to the deformation, of a particle that before the deformation was situated at point P having \mathbf{r} as position vector with respect to some arbitrary origin. A neighboring point Q, whose position vector was $\mathbf{r} + d\mathbf{r}$ before the deformation, will suffer a displacement $\mathbf{s}(\mathbf{r} + d\mathbf{r})$ which will differ from $\mathbf{s}(\mathbf{r})$ by the quantity

$$d\mathbf{s} = d\mathbf{r} \cdot \nabla \mathbf{s}$$

The hypothesis of small deformations means that $d\mathbf{s}$, the change in the displacement vector when we go from P to the neighboring point Q, is very small compared to $d\mathbf{r}$, the position vector of Q relative to P. Consequently, the scalar components of the dyadic $\nabla \mathbf{s}$ are all very small compared to unity. The geometrical meaning of the dyadic $\nabla \mathbf{s}$ is obtained by separating it into its symmetric part $\mathbf{S} = \frac{1}{2}(\nabla \mathbf{s} + \mathbf{s} \nabla)$ and its antisymmetric part $\mathbf{R} = -\frac{1}{2}\mathbf{1} \times (\nabla \times \mathbf{s})$, where $\mathbf{1}$ is the unity dyadic. The antisymmetric part is interpreted as follows: if at some point M the symmetric part vanishes, then we have for the neighborhood of M the relation

$$d\mathbf{s} = d\mathbf{r} \cdot \mathbf{R}_M = \boldsymbol{\omega}_M \times d\mathbf{r}$$

where $\boldsymbol{\omega}_M = \frac{1}{2}(\nabla \times \mathbf{s})_M$ is an infinitesimal vector. This means that the neighborhood of point M undergoes an infinitesimal rigid rotation, without

any change in shape or size. Consequently, the deformation is represented by the symmetric part \mathbf{S} , which is called the *strain dyadic*.

In a Cartesian orthonormal basis, in which we have $\mathbf{r} = \sum_{i=1}^3 x_i \mathbf{a}_i$, we write $\mathbf{s} = \sum_{j=1}^3 s_j \mathbf{a}_j$, and obtain

$$\mathbf{S} = \sum_{i,j=1}^3 \mathbf{a}_i \mathbf{a}_j S_{ij}$$

where $S_{ij} = \frac{1}{2} \left[\frac{\partial}{\partial x_i} s_j + \frac{\partial}{\partial x_j} s_i \right]$. The diagonal components S_{11} , S_{22} , and S_{33} are the coefficients of linear extension in the directions \mathbf{a}_1 , \mathbf{a}_2 , and \mathbf{a}_3 , respectively, while the non diagonal components $S_{12} = S_{21}$, $S_{13} = S_{31}$, and $S_{23} = S_{32}$ are called shear strains. For instance, $2S_{12}$ is the change in the angle of the dihedron formed by the planes that before the deformation were respectively normal to the directions \mathbf{a}_1 and \mathbf{a}_2 . The shear strains are not essential for the complete representation of a deformation since they can be made to vanish by expressing \mathbf{S} in the basis of its principal axes.

If an infinitesimal element of the body occupies the volume dV before the deformation and the volume dV' after, the relative increase of volume, or volumetric dilatation, is given by

$$\frac{dV' - dV}{dV} = S_{11} + S_{22} + S_{33} = |\mathbf{S}| = \nabla \cdot \mathbf{s}$$

The forces applied to a finite deformable body are either body forces acting on every volume element dV and represented by the notation $dV \mathbf{F}$ or $dV \rho \mathbf{K}$, where \mathbf{F} is the force per unit volume, \mathbf{K} is the force per unit mass, and ρ is the density, or surface forces acting on every element dS of the bounding surface and represented by $dS \mathbf{T}$, where \mathbf{T} is the surface stress, or surface force per unit area. The effect of these applied forces is transmitted throughout the body, so that through any surface element inside the body, there is a force exerted by the matter on one side of the element upon the matter on the other side. Such forces are called internal stresses and are defined as follows: let dS be a surface element completely inside the body, and let us choose arbitrarily the positive sense of the normal \mathbf{n} to this surface element; this defines for dS a positive side, the one containing \mathbf{n} , and a negative side. Then \mathbf{T}_n ,

the stress vector on the positive side of dS is defined as a vector such that $dS \mathbf{T}_n$ is the surface force on the positive side of dS —i.e., the resultant of all the forces exerted through dS by the matter on the positive side of dS upon the matter on the negative side. In general there is a normal component $\mathbf{T}_n \cdot \mathbf{n}$, which is a pressure or a traction depending upon whether the sign of $\mathbf{T}_n \cdot \mathbf{n}$ is negative or positive, and a tangent component $\mathbf{n} \times \mathbf{T}_n \times \mathbf{n}$ called the shear stress. The value of the stress vector \mathbf{T}_n depends upon the orientation of the normal \mathbf{n} , so that we can characterize the state of stress at a point by defining the *stress dyadic* \mathbf{T} through the relation

$$\mathbf{T}_n = \mathbf{n} \cdot \mathbf{T}$$

The mechanical equilibrium conditions applied to an arbitrary volume V , bounded by the closed surface S , and completely inside the deformable body give

$$\int_V dV \mathbf{F} + \int_S dS \mathbf{n} \cdot \mathbf{T} = 0$$

and

$$\int_V dV \mathbf{r} \times \mathbf{F} + \int_S dS \mathbf{r} \times (\mathbf{n} \cdot \mathbf{T}) = 0$$

By the use of the divergence theorem, the first condition gives the equation

$$\nabla \cdot \mathbf{T} + \mathbf{F} = 0$$

at any point inside the body, and the second condition implies that \mathbf{T} is a symmetric dyadic. On the external surface of the body, we have usually to fulfill the boundary condition

$$\mathbf{n} \cdot \mathbf{T} = \mathbf{T}$$

where \mathbf{T} is the applied external force per unit area. Other boundary conditions can also be met, such that the value of the displacement be prescribed.

For infinitesimal deformations, we assume that the relation between strain and stress is expressed by Hooke's law: the deformation is proportional to the applied force. For isotropic bodies, this linear relation is

$$\mathbf{S} = \frac{1}{E} [(1 + \nu) \mathbf{T} - \nu |\mathbf{T}| \mathbf{1}]$$

where E is Young's modulus and ν is Poisson's ratio. These two elastic constants can be defined by considering the stretching of a cylindrical bar by normal traction forces uniformly distributed on the end sections; then we have

Young's modulus :

$$\frac{\text{Normal traction force per unit cross sectional area}}{\text{Relative longitudinal extension}}$$

and

$$\text{Poisson's ratio} = \frac{\text{Relative lateral contraction}}{\text{Relative longitudinal extension}}$$

We can also write

$$\mathbf{T} = 2\mu \mathbf{S} + \lambda |\mathbf{S}| \mathbf{1}$$

where $\mu = E/2(1 + \nu)$ and $\lambda = \nu E/[(1 + \nu)(1 - 2\nu)]$ are Lamé's constants. μ is the rigidity modulus, the only constant necessary when the volumetric dilatation vanishes everywhere.

Substituting the preceding relation into the equilibrium equations, we transform them into

$$2\mu \nabla \cdot \mathbf{S} + \lambda \nabla |\mathbf{S}| + \mathbf{F} = 0 \text{ inside the body}$$

and

$$2\mu \mathbf{n} \cdot \mathbf{S} + \lambda \mathbf{n} |\mathbf{S}| = \mathbf{T} \text{ on the bounding surface.}$$

These vector relations are not sufficient for the complete determination of the symmetric dyadic \mathbf{S} . To insure that a solution of the above equations correspond to a possible displacement vector \mathbf{s} , we must be able to integrate the relation

$$\mathbf{S} = \frac{1}{2} (\nabla \mathbf{s} + \mathbf{s} \nabla)$$

i.e., from a given expression for \mathbf{S} obtain the value of \mathbf{s} . From the vanishing of the curl of a gradient, it is easily seen that this integrability condition, also called the compatibility equation, is

$$\nabla \times \mathbf{S} \times \nabla = 0$$

By elimination of the vector products, we obtain the equivalent form

$$\nabla \nabla \cdot \mathbf{S} + \nabla \cdot \mathbf{S} \nabla - \nabla \nabla |\mathbf{S}| - \nabla \cdot \nabla \mathbf{S} = 0$$

Using the stress-strain relation and the equilibrium conditions, we obtain the Beltrami-Michell form of the compatibility equation:

$$\nabla \cdot \nabla \mathbf{T} + \frac{1}{1 + \nu} \nabla \nabla |\mathbf{T}| = - \frac{\nu}{1 - \nu} \nabla \cdot \mathbf{F} \mathbf{1} - (\nabla \mathbf{F} + \mathbf{F} \nabla)$$

Finally, by expressing the strain dyadic in terms of the displacement vector, we obtain Navier's form of the equilibrium equations:

$$\mu \nabla \cdot \nabla \mathbf{s} + (\lambda + \mu) \nabla \nabla \cdot \mathbf{s} + \mathbf{F} = 0 \text{ inside the body}$$

and

$$\lambda \mathbf{n} \nabla \cdot \mathbf{s} + 2\mu \mathbf{n} \cdot \nabla \mathbf{s} + \mu \mathbf{n} \times (\nabla \times \mathbf{s}) = \mathbf{T}$$

on the bounding surface.

Dealing here directly with the displacement vector, there is no need of considering the compatibility equation.

The propagation equation for elastic disturbances is obtained by adding the inertia force to the body force. We get then

$$\mu \nabla \cdot \nabla \mathbf{s} + (\lambda + \mu) \nabla \nabla \cdot \mathbf{s} + \rho \mathbf{K} = \rho \frac{\partial^2 \mathbf{s}}{\partial t^2}$$

inside the body.

The stress-strain relation and the boundary conditions are not affected, but we generally have to take into account initial conditions.

The energy density u , or energy per unit volume, is given by

$$u = \frac{1}{2} \mathbf{S} : \mathbf{T} + \frac{1}{2} \rho \frac{\partial \mathbf{s}}{\partial t} \cdot \frac{\partial \mathbf{s}}{\partial t}$$

where the first term is potential, or strain energy, and the second term is kinetic energy. The energy flux density vector

$$\mathbf{S} = \frac{\partial \mathbf{s}}{\partial t} \cdot \mathbf{T}$$

is a vector such that $d\mathbf{S} \cdot \mathbf{n}$ gives the quantity of energy that flows per unit time through the surface element $d\mathbf{S}$ in the positive direction of \mathbf{n} , the normal to $d\mathbf{S}$. At any point the energy continuity equation

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{S} - \rho \frac{\partial \mathbf{s}}{\partial t} \cdot \mathbf{K} = 0$$

expresses the conservation of mechanical energy.

GÉRARD NADIAU

References

- Godfrey, D. E. R., "Theoretical Elasticity and Plasticity," London, Thames and Hudson Co., 1959.
 Green, A. E., and Zerna, W., "Theoretical Elasticity," New York, Oxford University Press, 1945.
 Nadeau, G., "Introduction to Elasticity," New York, Holt, Rinehart and Winston, Inc., 1964.
 Pearson, C. E., "Theoretical Elasticity," Cambridge, Mass., Harvard University Press, 1959.
 Sokolnikoff, I. S., "Mathematical Theory of Elasticity," New York, McGraw-Hill Book Co., Inc., 1956.

Cross-references: POLYMER PHYSICS, VECTOR PHYSICS, VISCOELASTICITY.

ELECTRIC POWER GENERATION

About 95 per cent of the electric power produced in this country is by 3-phase generators. It is transmitted and distributed this way. Advantages of 3-phase generators lie in economy of apparatus, lower transmission losses, inherent starting torque for polyphase motors, and constant running torque for balanced loading. A generator is built with axial slots for armature coils in a stationary hollow cylindrical iron core called the stator. The windings are placed in the slots so that when carrying current they produce a chosen even number of alternate magnetic poles. The coils over each magnetic pole are grouped in 3 equal bands to give a 3-phase balanced system of terminal voltages.

An inner rotor has coils which carry direct current to give the same number of alternate magnetic poles as on the stator. Rotor current strength is controlled by a rheostat or voltage from a dc generator. Voltages are produced in the stator windings by flux cutting as the rotor magnetic flux sweeps by them, and currents flow when the

generator terminals are connected to a 3-phase load impedance. The 3-phase stator line voltages are equal in magnitude and 120 electrical degrees apart in time sequence. So also are the line currents for a balanced 3-phase load. Generator voltages are of the order of 12 000 to 30 000 volts for large machines.

Generator frequency is the product of the pairs of magnetic poles and the speed in revolutions per second. At 60 cycles, a 2-pole generator runs at 3600 rpm and a 6-pole generator at 1200 rpm. The maximum speed of 3600 rpm has been increasingly adopted even for very large machines because high speed means decreased size and weight for a given kilowatt rating and better steam-turbine performance. Waterwheels and water turbines show best characteristics at much lower speeds—roughly a range of 100 to 600 rpm. Sixty cycles is the prevailing frequency in this country for public utility power generation. Because of weight and space limitations, 400 cycles is found in the aircraft industry. Europe is basically on 50 cycles.

In the large central station steam power plants, single generators may reach, or go somewhat beyond, 250 000 kW in rating. Some "units" (so-called) will have ratings up to a million kilowatts. Very large units are usually 2 separate single-shaft turbine-generator sets with one or two turbines on each shaft. One turbine takes steam at high pressure and the others at intermediate or low pressure.

Direct-current generators are built with their dc magnetic poles on the stator. Armature conductors on the rotor have ac voltages induced in them as they are rotated; the same principle of flux cutting holds as before. An automatic mechanical switching device, called a commutator, is placed on the shaft. It carries fixed brushes, and with its many insulated copper bars connected to the armature coils, it inverts every other alternation of the voltage to give unidirectional, or dc, voltage at the 2 armature terminals. It is the commutator that requires the rotor to be the armature so that coils and their switching arrangement always move exactly together. Direct-current generators are generally limited to several thousand kilowatts, and their application lies mainly in industrial plants.

Generators in stations of electric utility companies are driven by steam turbines, water wheels, or water turbines. The ratio of steam to hydro-power is 4 to 1. Industrial production of steam plus hydro is about one-twelfth that of utility steam alone. The installed generating capacity in this country is over 210 million kW, about one-third of the world total. The United States produces about four times the energy in kilowatt-hours as the rest of the world.

In addition to power generation by steam and hydro, electric power also is obtained from gasoline- and diesel-driven motor-generator sets; from batteries, fuel cells, and solar cells; by thermoelectricity and photoelectricity; and by wind motion. Most of these are described elsewhere in this volume. Power production by these

last methods is small compared with steam and water power. Windmills are used to charge batteries, but use of wave motion or tides has never proved feasible.

The only fuels of consequence used by steam plants are coal, oil, and gas. Coal has predominated in the eastern part of the United States, with oil and natural gas used in the western states. Burners for oil or gas plants may be adaptable for either type of fuel, or they can be interchanged quickly. In some instances, provision is also made for coal burning.

Nuclear steam plants have now demonstrated safety, economy, reliability, and excellent performance. They are competitive with conventional plants and indicate better capabilities. They conserve natural resources, give less air pollution, and in small sizes have applications not possible with other forms. There are now over 1 million kW installed in civilian reactor plants, and by 1980, 40 million kW are expected, about 17 per cent of the total required capacity in the United States. Expansion here and in Europe is rapidly advancing, and plans indicate 1 million kW in a single plant.

Geothermal production of electric power uses natural steam obtained from the earth through steam wells and piped to turbines. Italy produces about one-third of a million kW in this manner and New Zealand has slightly less installed capacity of this kind. The only U.S. installation is about 30 000 kW on the West Coast. Geothermal power is limited. Temperatures and pressures are low, but there is a lower capital investment and absence of fuel cost. A serious problem lies in elimination of contamination in the steam.

Major problems today include improvement of over-all characteristics and economy of existing apparatus, development of materials to withstand increasing temperatures and stresses; the disposal of combustion products and radioactive wastes, and finding new and enlarged power applications. Some envision the elimination of the conventional plant with its rotating machinery, associated equipment, noise, and maintenance.

Among new methods currently studied for power generation are the magnetohydrodynamic (MHD) generator, fuel cell, thermionic converter, thermoelectric generator, and fusion reactor. None are yet practical and economic for a utility company even though there have been a few minor applications. Of this group, the first appears to be most promising.

The limited efficiency of steam turbines imposed by the thermodynamic properties of steam has stimulated the development of methods to convert heat directly into electricity. The MHD generator is one in which a thermally ionized gas is forced at high temperature, pressure, and velocity through a duct situated in a transverse magnetic field. An induced voltage appears in the third mutually perpendicular direction (the Hall effect), and this voltage may be tapped by electrodes within the duct (see MAGNETO-FLUID-MECHANICS).

If the exhaust gas from the MHD generator is used to heat steam for a conventional generator, a

larger portion of the thermal spectrum will be utilized and the system efficiency may be raised from the present 40 per cent to possibly 50 or 55 per cent. Heat for the system may come from the use of fossil fuel or nuclear reactors.

B. L. ROBERTSON

References

Robertson, B. L., and Black, L. J., "Electric Circuits and Machines," Second edition, Princeton, N.J., D. Van Nostrand Co., 1957 (three-phase circuits, machine construction and performance).

Federal Power Commission yearly reports (power production and utilization in the United States).

Cross-references: BATTERIES, ELECTRICITY, MAGNETO-FLUID-MECHANICS, NUCLEAR REACTORS, PHOTOELECTRICITY, THERMOELECTRICITY.

ELECTRIC PROPULSION

Electric propulsion is a form of rocket propulsion in which electric power, generated on board the propelled vehicle, is used to eject propellant rearward at high velocity to produce thrust. Electric propulsion systems can be considered to be made up of two major components: (1) the *electric power generation system*, which converts power from a basic power source (such as a nuclear reactor or the sun) into electric power, and (2) the *thruster*, which uses this electric power to produce thrust by ejecting the propellant.

The primary potential advantage of electric rockets over chemical rockets or solid-core nuclear rockets is that much higher propellant ejection velocities can be attained. Higher ejection velocities, in accordance with Newton's law, produce higher thrust per unit mass of propellant, so that the total mass of propellant needed for space missions can be greatly reduced. The mass of the required electric power generation equipment is appreciable, however, so that some of the saving in propellant mass is offset by the mass of the power generation system. The net mass saving possible using electric propulsion, therefore depends strongly on the difficulty and duration of the mission as well as on the performance parameters of the system.

One of the most important of these performance parameters is the propulsion-system specific mass α , which is defined as

$$\alpha = \frac{m_{\text{ps}}}{P_j} \frac{kg}{kW} \quad (1)$$

where m_{ps} is the total propulsion system mass (in kilograms) and P_j is the jet power produced (in kilowatts). If this parameter is less than about 15 kg/kW, electric propulsion systems can be employed to advantage over nuclear or chemical rockets for most unmanned interplanetary exploration missions. For such missions, typical required power levels range from several hundred kilowatts to a megawatt, to propel vehicles having

initial mass in earth orbit in the range of 10 000 to 100 000 kg.

If α is less than about 5 kg/kW, electric propulsion is superior to nuclear rockets, with regard to required initial weight and trip time, for manned expeditions to the near planets.¹ For these missions, power levels of several megawatts will be needed for vehicle weights (in orbit) of the range of 100 000 to 1 000 000 kg. In other possible applications, such as providing small amounts of thrust for attitude control or orbit control of satellites, the specific mass is less important, since the required electric power is small and can usually be obtained from the power supply used by the other on-board equipment.

Most of the mass of an electric propulsion system resides in the electric power generation system; however, the performance of the other major component, the thruster, is of equal importance in determining the over-all specific mass. The most important parameter for the thruster is the efficiency η with which the electric power is converted into jet power. If this efficiency is low, the required electric power, and therefore the power-plant mass, is correspondingly high.

Another important parameter for the thruster (as for all rockets) is the specific impulse I . This parameter is defined as the thrust F produced per unit weight flow of propellant:

$$I = \frac{F}{\dot{m}_p g_0} \quad \text{sec} \quad (2)$$

where \dot{m}_p is the mass flow rate of propellant and g_0 is the acceleration of gravity at the earth's surface (9.8 m/sec²), which relates mass to weight. The relation of thrust, specific impulse, and propellant ejection velocity is

$$F = \dot{m}_p g_0 I = \dot{m}_p v_j \quad \text{newtons} \quad (3)$$

where v_j is the mean propellant ejection velocity (more commonly called *effective jet velocity*). The first and last terms in Eq. (3) express Newton's law that force is equal to the time rate of change of momentum. The last two terms show that specific impulse is directly proportional to effective jet velocity.

The *jet power* is the time rate of change of jet kinetic energy, or

$$P_j = \frac{1}{2} \dot{m}_p v_j^2 = \frac{1}{2} F v_j = \frac{1}{2} g_0 I F \quad \text{newton-m/sec}$$

or, in kilowatts,

$$P_j = \frac{g_0 I F}{2000} \quad \text{kW} \quad (4)$$

For constant thrust and jet velocity, the total propellant mass m_p needed for a mission can be written [from Eq. (3)] as

$$m_p = \frac{F t}{g_0 I} \quad \text{kg} \quad (5)$$

where t is the total propulsion time and $F t$ is the total impulse required for the mission. From

Eq. (1) and (4), the propulsion system mass can be written:

$$m_{ps} = \alpha P_j = \frac{\alpha g_0 I F}{2000} \quad \text{kg} \quad (6)$$

These equations show that, although propellant mass can be reduced indefinitely by increasing the specific impulse [Eq. (5)], the power required (and therefore the power-plant mass) is increased when this is done [Eq. (6)]. It is, therefore, desirable to use that value of specific impulse for which the *sum* of the masses of propellant and propulsion system is lowest. This optimum specific impulse will yield the least total mass for the mission, or the highest payload mass, for a given total mass. For lunar and interplanetary missions and for specific weights likely to be obtained, calculations show that the optimum specific impulses range from about 1500 to 15 000 seconds (corresponding to jet velocities of about 15 to 150 km/sec). These specific impulses compare with values of about 450 seconds that are typical for high-energy chemical rockets and about 900 seconds that may be possible with solid-core nuclear rockets.

Another characteristic feature of electric propulsion systems is the very low thrust generated in comparison with chemical or nuclear rockets. This can be seen from Eq. (6) which can be written:

$$\frac{F}{m_{ps} g_0} = \frac{2000}{\alpha I g_0^2} \quad (7)$$

For a specific weight α of 10 kg/kW, and a specific impulse I of 5000 seconds, Eq. (7) yields a thrust-to-weight ratio of about 4×10^{-4} . This very low value results partly from the higher specific impulse typical of electric rockets, but mostly from the specific weight, which is of the order of 1000 or more times higher than that obtainable with solid-core nuclear rockets or chemical rockets. The low thrust-weight ratio means that electric propulsion systems cannot be used for launching from planetary surfaces. They are best suited for propelling vehicles between orbits about the planets or between orbits about the earth and the moon.

Because the thrust-weight ratio is so low, electric rockets must operate for much longer periods of time (of the order of 1000 times longer) than chemical or nuclear rockets to produce the same total impulse. Typically, for interplanetary missions to the near and far planets, these required operating times range from many months to several years. The removal of limitations on jet velocity, therefore, is obtained at the expense of greatly increased propulsion system weight and required operating lifetime.

Power Generation Systems. The need for low specific weight dominates the selection of suitable methods for generating electric power for primary propulsion of space vehicles. The requirement that power be generated with very little consumption of mass dictates that either nuclear or solar energy must be used as the basic energy source.

Among the possible methods of converting this energy into electric power, the most direct are photovoltaic solar cells and radioisotope cells. Considerable progress has been made in reducing the thickness, and hence the weight of photovoltaic solar cells;² eventual achievement of a specific mass near 5 kg/kW appears possible. A lightweight radioisotope cell, in the range of 1 kg/kW, has been proposed and analyzed³ but not yet demonstrated. This cell is basically a very high-voltage, low-current device, which converts a large fraction of the kinetic energy of the isotope decay particles directly into electric power. This system matches the requirements with respect to voltage and current, of colloidal-particle thrusters (see "Thrusters," p. 187).

Somewhat less direct in energy conversion are systems that use thermionic cells

$$\begin{pmatrix} \text{nuclear} & \text{--} & \text{heat} & \text{--} & \text{electricity} \\ \text{solar} & & & & \end{pmatrix}$$

A nuclear reactor or solar concentrator is used to heat a suitable material (such as tungsten) to temperatures high enough to produce thermal emission of electrons. These electrons traverse a gap to a cooled collector electrode, thereby producing electric power at a potential of the order of 1 volt. Many thousands of these thermionic cells must be connected in series-parallel combinations to achieve the required power levels and voltages. Also, to produce useful power densities, emitter temperatures must be in the range 1500 to 2000 K. Conversion efficiencies (heat into electric power) of 15 to 20 per cent are possible. The remaining 80 to 85 per cent of the thermal power must be radiated into space. The collector electrodes, where this waste heat appears, are, in the lighter-weight configurations, buried within the nuclear reactor or the heat absorber of the solar collector, so that a heat-transfer fluid must be pumped past the collector to pick up the waste heat and carry it to a radiator. In order that the radiator be of adequately low size and weight, it must operate at temperatures of about 1000 K or

higher. Analyses for a complete nuclear thermionic system yield specific masses of the order of 4 to 10 kg/kW, but numerous severe performance, design, and engineering problems remain to be solved before such systems can be developed to mission status.¹

Still more indirect, in the conversion of energy, are the turbopropeller systems

$$\begin{pmatrix} \text{solar} & \text{--} & \text{heat} & \text{--} & \text{mechanical} & \text{--} & \text{electric} \\ \text{nuclear} & & & & & & \end{pmatrix}$$

For these, as well as the thermionic systems, the nuclear reactor appears to be a better basic energy source than the sun, because it provides a more compact and versatile system, suitable for operation in shaded regions and at any distance from the sun.

A nuclear turbopropeller system for electric propulsion (as illustrated in Fig. 1) is basically a lightweight adaptation to space conditions of ground-based nuclear power stations.⁵ The chief differences result (as for the thermionic systems) from the lack of means other than radiation to eliminate the waste heat resulting from inescapable conversion inefficiencies. To produce specific weight below 10 kg/kW the waste-heat radiator must operate at temperature above 900 K, which in turn requires that the nuclear reactor operate at temperatures in excess of 1200 K.

The most suitable working fluid, at these temperatures, is potassium if a liquid-vapor thermodynamic cycle (Rankine cycle) is used. In a single-loop version of this cycle, the liquid metal is vaporized in the nuclear reactor; the resulting vapor drives the turbine, which in turn drives the generator to produce electric power. The vapor passes from the turbine through the radiator, where it is recondensed, and the liquid is then recirculated through the reactor. A major problem is to develop materials with adequate corrosion resistance during long periods of high-temperature operation with alkali liquid metals. As illustrated in Fig. 1, the radiator is the largest and heaviest part of the system. The necessarily

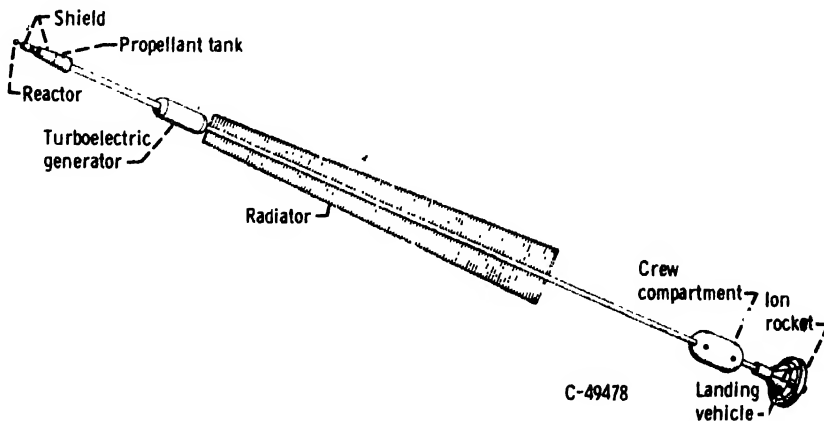


FIG. 1. Conceptual design of space vehicle for manned Mars mission. Nuclear turbopropeller propulsion system.

large exposed area is vulnerable to micrometeoroid penetration. Until more information is available on the flux and distribution of these particles in space, the precise wall thickness (and weight) needed to provide negligible penetration probability cannot be accurately estimated. Current evidence, however, suggests that the micrometeoroid hazard is less severe than previously estimated. Of all power generating systems suitable for electric propulsion, the Rankine-cycle, liquid-metal nuclear turboelectric system is nearest the stage of development for space missions.

Thrusters. A large number of methods are possible to eject propellant by use of electric power. These are generally divided into three categories: (1) *electrostatic thrusters*, in which atoms (or heavier particles) are electrically charged and then accelerated rearward by means of an electrostatic field; (2) *plasma thrusters*, in which the propellant is made into an electrically conducting gas and accelerated rearward by application of electromagnetic forces; (3) *electro-thermal thrusters*, which use the electric power to heat the propellant, and then accelerate it rearward by thermal expansion through a nozzle.

Electrostatic thrusters that accelerate atomic ions (ion rockets) have received the most research and development attention, as a result of early demonstrations of good efficiencies in the range of specific impulses needed for major space missions. Typical of these ion rockets is an electron-bombardment thruster (such as that shown in Fig. 2) which uses mercury vapor as propellant.⁶ The propellant atoms are ionized by collision with electrons emitted by the cathode

and attracted toward the anode. A weak axial magnetic field is maintained in the ionization chamber to make the electrons spiral around on their way to the anode, thereby increasing their path length and their probability of colliding with propellant atoms. The resulting positive ions are extracted through a screen grid by means of an accelerating grid that is maintained at the proper voltage difference (usually several thousand volts) to produce the desired ejection velocity (specific impulse). A second electron emitter (not shown) is placed adjacent to the ion beam, downstream of the accelerator, to neutralize both the ion space charge and the net current leaving the thruster. Experimental efficiencies in converting electric power into jet power range from 60 to 80 per cent at specific impulses in the range 4000 to 9000 seconds. Thrusters in sizes up to 50 cm in diameter, with jet powers near 30 kW have been successfully operated.²

Other ion thrusters, using contact ionization of cesium atoms on hot tungsten to produce the ions (rather than electron bombardment), have achieved somewhat lower performance. In these thrusters, cesium vapor is passed through porous tungsten, which must be heated to about 1500 K to evaporate enough cesium ions from the ionizer surface. The high work function of tungsten and the low ionization potential of cesium make these two substances the most promising for contact ionization thrusters.

Atomic-ion thrusters tend to become less efficient at low ejection velocities (low specific impulse), because a certain fixed amount of energy is needed to ionize the propellant atoms. As the ejection velocity decreases, the jet power

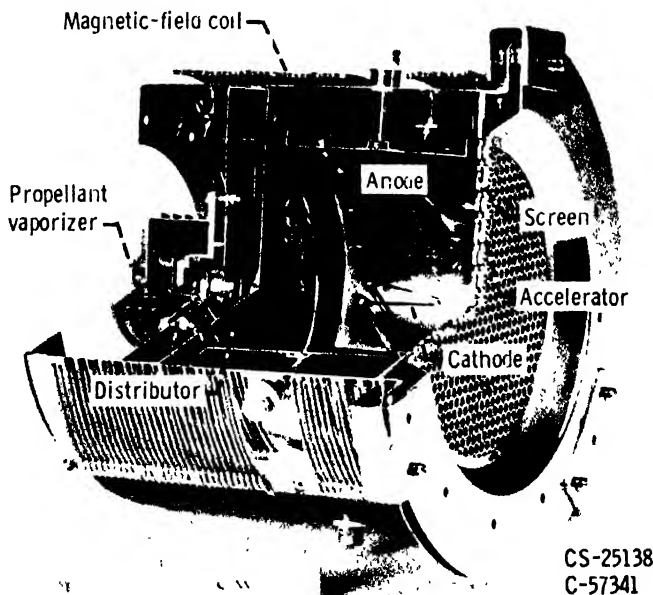


FIG. 2. Cutaway photograph of electron-bombardment ion thruster. With about 1 kW of power, this unit produces a thrust of 2.5 grams at a specific impulse of 5000 sec.

approaches the power required for ionization, and the efficiency decreases. A possible way to increase the efficiency is to increase the mass of each charged particle so that its kinetic energy, at a given jet velocity, is higher. This approach leads to use of colloidal particles in place of atomic ions. Because of the much higher mass per unit charge, voltages in the hundreds of kilovolts are needed to produce the desired jet velocities. Research is underway to find ways of efficiently generating, charging, and accelerating such particles.⁸

Although the efficiencies attainable with electrostatic thrusters are high, there remains a limitation which, although not crucial, is undesirable, namely, a low thrust (or power) per unit beam area, due to limitations on ion beam current density. These limitations result from two sources: (1) space charge and (2) accelerator electrode erosion. The space-charge limited current is determined by the accelerating voltage and the distance between accelerator electrode and ion source. The voltage, in turn, is approximately fixed by the desired specific impulse, and the accelerator spacing is limited by electrical breakdown and thermal warping. The erosion limitation appears to be even more restrictive on thrust per unit area than space charge.⁸ As the current density is increased, there is greater impingement of ions on the accelerator electrode. For an accelerator lifetime of the order of 1 year, estimates indicate a limit for thrust per unit beam area of about 2 newtons (0.2 kg) per square meter (about 50 kW/m²) at a specific impulse of 5000 seconds. Higher values are allowable as specific impulse increases.

Plasma thrusters, which operate on the principle of accelerating an electrically conducting but neutral gas (plasma), are not subject to the space-charge limitation. Furthermore, some require no electrode contact with the propellant stream. Consequently, a higher thrust per unit area, with adequate operating lifetime, may be possible with plasma thrusters than with ion thrusters. A variety of plasma thrusters are currently being investigated,⁹ but so far the efficiencies have been much lower than those of the ion thrusters.

Electrothermal thrusters are primarily of two types—the electric-arc jet and the electrically powered hydrogen heater (also called resistojets). The hydrogen heater is limited to specific impulses less than about 1000 seconds, because of the limitation on the wall temperature of the heater. High efficiencies, however, have been achieved.¹⁰ The arc jet, which heats the propellant by means of continuous electric discharge as the propellant flows by, can achieve somewhat higher specific impulses (up to about 2000 seconds), but the efficiency is generally less than 50 per cent, due to losses involved in dissociation and ionization of the propellant atoms, and losses to the walls of the arc chamber and nozzle. Because of the lower specific impulse range, electrothermal thrusters are not useful for interplanetary missions, but may be used for more limited applica-

tions such as satellite orientation control and orbit correction.

Some hybrid thrusters, which are combinations of the three major categories, have also been studied. One of these is an electrostatic ion accelerator in which the electrons are held by a magnetic field in the accelerating region while the ions pass through.⁹ In this device, the space-charge limitation on ion current density is removed, but as yet the efficiency has not been good. Another device is an arc-jet that is operated at very low gas density but high arc current. In this device, in addition to the heating, there is an electromagnetic acceleration due to interaction of the high arc current with its own, or an externally imposed, magnetic field. In this way, specific impulses of as much as 12 000 seconds have been attained, with efficiencies that are high, but not as high as those obtained with atomic-ion thrusters.¹¹

History and Status. The possibility of reducing propellant consumption by ejecting the propellant electrically at high velocities was recognized by early space flight and rocket pioneers, such as Goddard¹³ and Oberth,¹² but the practical feasibility of such propulsion systems was not demonstrated. With the advent of nuclear-electric power and large rockets during and after World War II, more interest in electric propulsion was aroused, and between 1946 and 1956, a number of preliminary analyses of nuclear and solar electric systems were published.^{14,15} The use of cesium-tungsten electrostatic thrusters was first proposed by Stuhlinger,¹⁵ and early experimental work in the United States, beginning in 1957, was concentrated on this approach.¹⁶⁻¹⁹

A systematic, comparative study of the applicability of electric, nuclear, and chemical propulsion to future space missions, together with an engineering study of large electric power systems for space use, were completed in 1957.²⁰ These and similar studies led to the initiation of major research programs in electric propulsion and power generation at U.S. government and industrial laboratories. Since, 1957, under the stimulus of a major national space program, these studies have expanded greatly and have covered a wide spectrum of promising methods.

At the same time, many low-thrust trajectory studies and mission analyses²¹⁻²⁵ have further clarified the role of electric propulsion in future space missions.

Numerous facilities for testing electric thrusters and components of power generation systems are now available. Vacuum tanks at the NASA Lewis Research Center, for example, permit testing of complete electric thruster systems up to the megawatt power range, and facilities under construction will permit vacuum testing of complete nuclear-electric power and propulsion systems in the sizes needed for unmanned and manned interplanetary travel. If the current research and development programs prove successful, electric propulsion systems suitable for unmanned interplanetary exploration could be

available in the early or middle 1970's, and systems for manned interplanetary exploration of the near planets by the late 1970's. The major problems are those of producing systems with adequate operating lifetime and reliability while maintaining low weight and high efficiency.

W. E. MOECKEL

References

1. Moeckel, W. E., "Electric Propulsion," *Science*, **142**, No. 3589, 172-178 (Oct. 11, 1963).
2. Shirland, F. A., Griffin, T. A., and Dierssen, G. H., "Thin Film CdS Front Wall Solar Cells," Paper 2566-62, ARS, 1962.
3. Mickelsen, W. R., and Low, C. A., Jr., "Potentials of Radioisotope Electrostatic Propulsion," *Astronautics Aerospace Eng.*, **1**, No. 9, 52-57 (October, 1963).
4. Bernatowicz, Daniel T., "A Parametric Study of the Thermionic Diode System for Large Nuclear-Electric Power-Plants in Space Vehicles," NASA TN D-1110, 1962.
5. English, Robert E., Slone, Henry O., Bernatowicz, Daniel T., Davison, Elmer H., and Lieblein, Seymour, "A 20,000-Kilowatt Nuclear Turboelectric Power Supply for Manned Space Vehicles," NASA MEMO 2-20-591, 1959.
6. Kaufman, Harold, R., "An Ion Rocket with an Electron-Bombardment Ion Source," NASA TN D-585, 1961.
7. Reader, P. D., and Mickelsen, W. R., "Experimental Systems Studies of Large Modules and Arrays of Electrostatic Thrusters," First Annual Meeting and Tech. Display of AIAA, Wash., D.C., June 28-July 2, 1964, NASA TM X-52019.
8. Mickelsen, W. R., and Kaufman, H. R., "Status of Electrostatic Thrusters for Space Propulsion," NASA TN D-2172, 1964.
9. Seikel, George R., "Generation of Thrust Electromagnetic Thrusters," Proc. NASA-Univ. Conf. on Sci. and Tech. of Space Exploration, Chicago (Ill.), Nov. 1-3, 1962, pp. 171-176, NASA SP-11, 1962.
10. Jack, John R., "NASA Research on Resistance Heated Hydrogen Jets, Advanced Prop. Concepts, Vol. 1, pp. 75-89, 90-92, New York, Gordon and Breach Science Pub., Inc., 1963.
11. Yaffee, Michael L., "Three Firms Developing Hybrid Thruster," *Airiation Week and Space Tech.*, **80**, No. 12, 65-69 (March 23, 1964).
12. Oberth, H., "Wege zur Raumschiffahrt," Munchen und Berlin, Verlag von Oldenbourg, 1929 (reprinted by Edwards Bros., Inc., 1945).
13. Lehman, Milton, "This High Man," New York, Farrar, Straus, & Co., 1963.
14. Shepherd, L. R., and Cleaver, A. V., "The Atomic Rocket," Pt. I, *J. Brit. Interplanet. Soc.*, **7**, 185-184 (1948); Pt. II, *ibid.*, **7**, 234-241 (1948); Pt. III, *ibid.*, **8**, 23-37 (1949); Pt. IV, *ibid.*, **8**, 59-70 (1949).
15. Stuhlinger, E., "Electrical Propulsion System for Space Ships with Nuclear Power Source," *J. Astronautics*, **2**, No. 4, 149-152 (1955); **3**, No. 1, 11-14 (1956); **3**, No. 2, 33-36 (1956).
16. Forrester, A. T., and Speiser, R. C., "Cesium-Ion Propulsion," *Astronautics*, **4**, No. 10, 34-35, 92, 94, 96-97 (October, 1959).
17. Childs, J. H., "Design of Ion Rockets and Test Facilities," Paper 59-103, Inst. Aero. Sci., Inc., 1959.
18. Mickelsen, William R., "Electric Propulsion for Space Flight," *Aerospace Eng.*, **19**, 6-11 (November, 1960).
19. Brewer, G. R., Etter, J. E., and Anderson, J. R., "Design and Performance of Small Model Ion Engines," Paper 1125-60, ARS, 1960.
20. Moeckel, W. E., Baldwin, L. V., English, R. E., Lubarsky, B., and Maslen, S. H., "Satellite and Space Propulsion Systems," NASA TN D-285, 1960. (Unclassified version of material presented at NACA Flight Propulsion Conference, November 22, 1957.)
21. Irving, J. H., and Blum, E. K., "Comparative Performance of Ballistic and Low-Thrust Vehicles for Flight to Mars," *Vistas Astron.*, **2**, 191-218 (1959).
22. Moeckel, W. E., "Fast Interplanetary Missions with Low-Thrust Propulsion Systems," NASA TR R-79, 1960.
23. Sauer, C. G., and Melbourne, W. G., "Optimum Earth-to-Mars Trip Trajectories Using Low-Thrust, Power-Limited Propulsion Systems," Rep. TR 32-376, Jet Prop. Lab., C.I.T., 1963.
24. Beale, Robert J., Speiser, Evelyn W., and Womack, James R., "The Electric Space Cruiser for High-Energy Missions," Rep. TR 32-404, Jet Prop. Lab., C.I.T., June 8, 1963.
25. Stearns, John W., "Electrical Propulsion Requirements for Planetary and Interplanetary Spacecraft," Rep. 32-403, Jet Prop. Lab., C.I.T., March 1, 1964.

Cross-references: ASTRODYNAMICS; ASTRONAUTICS; PHYSICS OF DYNAMICS; FLIGHT PROPULSION FUNDAMENTALS; IMPULSE AND MOMENTUM; MAGNETO-FLUID-MECHANICS; PHOTOELECTRICITY; PLASMA.

ELECTRICAL DISCHARGES IN GASES

Motion of Slow Electrons in Gases. Suppose that a swarm of electrons traverses a gas in which a uniform electric field X exists. In general the distribution of energy among the electrons will depend on the distance x which they have traveled in the field. However, provided x is sufficiently large, the energy distribution attains a steady value independent of x . In this steady state, the average rate of supply of energy to an electron from the field is equal to the average rate of loss of energy in collisions with gas molecules.

Many important quantities in this subject are related to $eX\lambda$, the energy gained by an electron of charge e in traveling the mean distance λ between two successive collisions with gas molecules. Since λ is inversely proportional to the gas density, the above quantity can be expressed in the form $X P_0$, where P_0 is the gas pressure reduced to some standard temperature.

The mean energy of an electron in the swarm, $\bar{\epsilon}$, is a function only of $X P_0$ for a particular gas. Figure 1 shows the form of this variation for a monatomic gas (He) and a diatomic gas (N₂). Here $k = \bar{\epsilon}/\bar{\epsilon}_k$, where $\bar{\epsilon}_k$ is the mean kinetic energy of a gas molecule at 15°C (0.037 eV). It is seen that the mean electron energy greatly

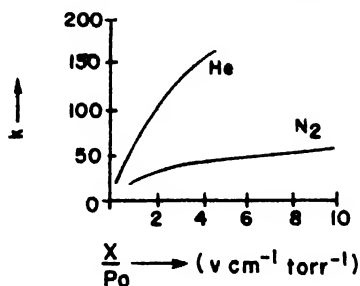


FIG. 1. Mean electron energy as a function of X/P_0 for He and N_2 .

exceeds the mean energy of a gas molecule even when X/P_0 is small. This is due to the inefficient energy exchange in collisions between electrons and gas molecules. If the collisions are elastic, it is readily shown that $\bar{\epsilon}$, the mean fractional energy lost by an electron in a collision, is $\sim 2m/M$ where m is electron mass and M is molecular mass. Clearly $f_e \ll 1$. At a given X/P_0 , $\bar{\epsilon}$ is generally lower in polyatomic than in monatomic gases. Owing to the possibility of inelastic collisions involving vibrational or rotational excitation of the molecule, $f \gg 2m/M$ in the former case. In the latter case, only electronic excitation of the atom can occur and this requires much higher energies in general.

In addition to their random motion, the electrons must obviously possess a superimposed drift motion in the direction of the applied field. Figure 2 shows the variation of the drift velocity W_e with X/P_0 ; normally W_e is small compared to the mean random speed of the electrons.

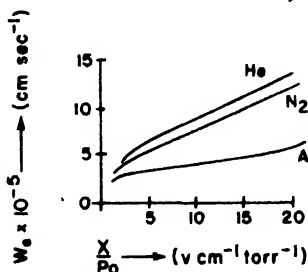


FIG. 2. Electron drift velocity as a function of X/P_0 for He, A and N_2 .

Ionization by Electron Collision. When the energy of an electron exceeds a certain critical value ϵ_i , ionization can occur at a collision with a gas molecule. As X/P_0 , and hence $\bar{\epsilon}$, is increased, an increasing fraction of electrons in the swarm will have energies exceeding ϵ_i . The size of the electron swarm will then increase with the distance x traveled in the field direction.

This growth is most conveniently studied under conditions where $\bar{\epsilon}$ is kept constant. This can be done by releasing electrons from the cathode

of a plane-parallel system and varying the electrode gap d and potential difference V in such a way that the electric field $X (= V/d)$ is fixed. It is then found that the electron current at the anode, i , increases exponentially with d or V . That is

$$i = i_0 \exp(\eta V) \quad (1)$$

where i_0 is the electron current released from the cathode and η is the electron ionization coefficient: this is defined as the average number of ionizing collisions made by an electron in moving through a 1-volt potential difference. η is also a function only of X/P_0 for a given gas (Fig. 3).

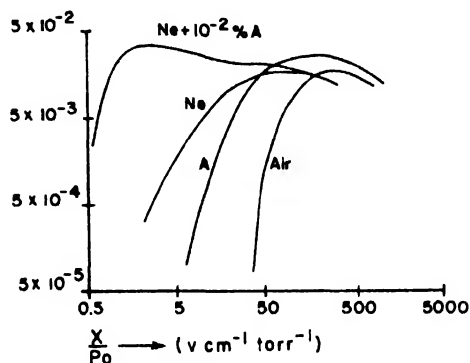


FIG. 3. Electron ionization coefficient as a function of X/P_0 for Air, Ne, A and $Ne + 10^{-2}\% A$.

It is important to note that the curve of η against X/P_0 passes through a maximum. The decrease in η at low X/P_0 is due to the increasing importance of excitation compared with ionization as X/P_0 decreases; since the excitation energy losses are larger in polyatomic than in monatomic gases, as remarked earlier, the decrease in η occurs more rapidly in air than in neon. The decrease in η at high X/P_0 (300 volts $\text{cm}^{-1} \text{ torr}^{-1}$), where excitation losses are comparatively unimportant, is due to the fact that an increasingly large fraction of the energy supplied from the field is used in maintaining the kinetic energy of the swarm.

The curve for the gas mixture $Ne + 10^{-2}\%$ per cent A is of great interest. Since the excitation potential of the most important metastable state of Ne (16.5 volts) exceeds the ionization potential of A (15.8 volts), the process $Ne^* + A \rightarrow Ne + A^+ + e$ can occur. This reaction has a very high probability, of the order unity per collision, and causes a great increase in η at low X/P_0 , above the value for pure neon, since the effective excitation energy losses are now considerably reduced. The double maximum in the curve of η vs X/P_0 arises from the fact that the direct and indirect ionization processes have their maximum efficiencies at different X/P_0 values (~ 70 and 2 volts $\text{cm}^{-1} \text{ torr}^{-1}$, respectively).

Secondary Ionization Processes. It is found that Eq. (1) no longer holds at larger values of

V_s ; i now increases more rapidly leading ultimately to spark breakdown. This is due to the occurrence of secondary ionization processes, in addition to ionization by collision between electrons and gas molecules. In general, the most important secondary process is the release of electrons from the cathode surface. If various simplifying assumptions are made, it can be shown that the ionization current is now given by:

$$i = \frac{i_0 \exp(\eta V)}{1 - \gamma[\exp(\eta V) - 1]} \quad (2)$$

where γ is a generalized secondary ionization coefficient. This is defined as the probability of a secondary electron being released from the cathode per positive ion arriving at the cathode. Included in γ are contributions to the secondary emission arising from radiation quanta and metastable molecules.

Since γ depends largely on the mean energies of the electrons and ions, it is, like η , a function only of X/P_0 though the function now depends on the nature of the cathode as well as on the gas.

Spark Breakdown. It is clear from the above equation that the ionization current tends to become very large as the potential difference across the gap approaches the value V_s given by:

$$\eta V_s = \log \left(1 + \frac{1}{\gamma} \right) \quad (3)$$

This is the condition for spark breakdown and can be best explained in the following manner.

Suppose that a primary electron current i_0 is released from the cathode when $V = V_s$. The electron current reaching the anode is then $i_0 \exp(\eta V_s)$. Hence, the positive ion current reaching the cathode due to the current i_0 is $i_0[\exp(\eta V_s) - 1]$. This will give rise to a secondary electron current of value $\gamma i_0[\exp(\eta V_s) - 1]$. If V_s is given by Eq. (3), then $\gamma[\exp(\eta V_s) - 1] = 1$ and the secondary current is equal to the original primary current i_0 . Hence it is clear that the process can continue even if the initiating current ceases. When V is less than V_s , however, the discharge current i is proportional to i_0 [Eq. (2)]. Thus $i = 0$ when $i_0 = 0$. It follows that $V = V_s$ marks the transition from a non-self-maintained to a self-maintained discharge. V_s is best defined as the potential difference required to maintain a small discharge current i when the primary current $i_0 = 0$. V_s is independent of i provided this is sufficiently small to avoid space charge distortion of the field.

Since η and γ are both functions only of X/P_0 and $X = V_s/d_s$ at breakdown, it follows from Eq. (3) that

$$V_s = F(P_0 d_s) \quad (4)$$

Thus, for a given gas and cathode material, the breakdown potential between large plane-parallel electrodes depends only on the product of the reduced gas pressure and electrode separation. This result, which is known as Paschen's law,

has been confirmed experimentally over a wide range of P_0 and d_s .

The variation of V_s with $P_0 d_s$ for a number of gases and cathode materials is shown in Fig. 4. It should be noted that the curves all exhibit a minimum; this corresponds to the maximum in the curve of η vs X/P_0 . It will be seen that the rise of V_s at high values of $P_0 d_s$ is most marked in air, less in pure Ne, and less still in the Ne + A mixture. This is readily understood by reference to the decrease in η at low X/P_0 ($= V_s/P_0 d_s$) in these gases (Fig. 3).

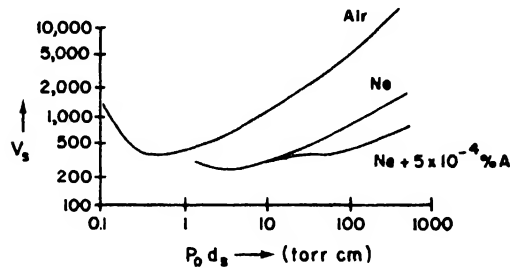


FIG. 4. Breakdown potential as a function of the product of the reduced gas pressure and electrode separation, $P_0 d_s$ for Air, Ne, and Ne + 0.0005% A with an iron cathode.

Time Lag of Spark Breakdown. If a potential difference $\geq V_s$ is suddenly applied to a discharge gap, a finite time elapses before the initial current i_0 has increased to a self-maintained discharge current $\sim 10^{-7}$ ampere/cm². This time lag consists of two parts. First of all, there is a statistical lag which arises from the fact that the primary and secondary ionization processes are both subject to statistical fluctuations. Thus, although $V = V_s$ where $\gamma[\exp(\eta V_s) - 1] = 1$ implies that on the average one electron leaving the cathode will give rise to one secondary electron, this may not happen in any particular case. Clearly the mean statistical lag t_s will decrease as the initial current is increased and it may be shown that

$$t_s = \frac{1}{PN_0} \quad (5)$$

where N_0 is the number of primary electrons leaving the cathode per second and P is the probability that any particular electron leads to breakdown. The latter quantity is zero at the sparking threshold V_s but increases rapidly for $V > V_s$. $P \approx 1$ provided $V > 1.25 V_s$.

The second component of the total time lag is the formative lag t_F . This can be regarded as the time that must elapse after the appearance of a suitable initiatory electron before the various ionization processes generate a self-maintained current of any given magnitude. This current can be chosen arbitrarily to specify breakdown of the gap and is generally taken to be $\sim 10^{-7}$ ampere/cm². Clearly t_F will depend on the relative importance of the various secondary mechanisms mentioned earlier; positive ion

transit times are typically $\sim 10^{-6}$ second, while the time lags involved in the contribution of radiation quanta and metastable molecules to γ are $\sim 10^{-8}$ and 10^{-3} second, respectively. The observed variation of t_F with overvoltage ΔV ($=V - V_*$) for various fixed values of X/P_0 in H_2 is shown in Fig. 5. Comparison with theory enables an estimate to be made of k , the relative contribution of photons at the cathode to the total γ . This ranges from 0.75 at $X/P_0 = 50$ to 0.50 at $X/P_0 = 300$ volts cm^{-1} torr $^{-1}$.

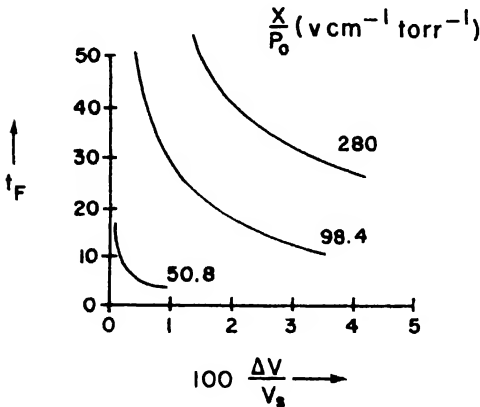


FIG. 5. Variation of formative time lag with overvoltage for various fixed values of X/P_0 in H_2 with copper electrodes

Glow Discharge. We have seen that any small current i can be maintained even in the absence of initiatory electrons when the potential difference between the electrodes reaches a value V_* given by Eq. (3). V_* is only independent of i when the latter is less than $\sim 1 \mu A$. At higher currents the space charge concentration becomes sufficient to cause X and hence η to vary across the gap, and ηV_* in Eq. (3) must be replaced by $\int \eta dV$. The static $V-i$ characteristic is normally negative since the field redistribution produced by the space charge effects increases the over-all ionization efficiency. Once breakdown has taken place, the current increases to a value determined by the impedance of the voltage supply.

If the current density is sufficiently small (< 0.1 ampere/cm 2), the cathode is not heated to a high enough temperature for thermionic emission to be a significant factor in the maintenance of the discharge. This regime is termed a glow discharge and the field variation across the gap in a long cylindrical tube is indicated in Fig. 6. We can distinguish five main regions here:

(1) The cathode fall, in which the field decreases from a high value at the cathode to approximately zero.

(2) The negative glow, in which ionization and excitation are due largely to fast electrons arriving from the cathode fall. The length of this region is normally controlled by the distance traveled by the electrons before their energy is reduced below the minimum required for excitation.

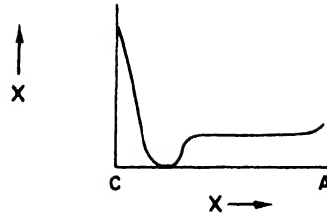


FIG. 6. Variation of axial field with distance from the cathode for a glow discharge in a long cylindrical tube.

(3) Faraday dark space. In many cases the ionization in the negative glow is so intense that the electron current here exceeds the total discharge current. A region is therefore required where electrons are lost by diffusion and not replenished by ionization; usually $X \leq 0$ here.

(4) The positive column, where X has a small constant value such that the corresponding electron energy distribution gives an ionization rate which just balances the loss of electrons and ions by radial diffusion to the walls.

(5) The anode fall, where X again increases.

Regions (1) and (2) are the most important regions of the discharge; the primary and secondary ionization by which the discharge is maintained take place here. The fall of potential across region (1), usually termed the cathode fall (V_*) is clearly an important parameter of the discharge.

It should be noted that the section extending from the negative glow to the anode has only a small field strength and small resultant space charge with $|n_i - n_e| \ll n_e$, where n_i and n_e are the ion and electron concentrations. This region is generally called a plasma. In many cases, the electrons here have a random motion which is large compared to their drift motion in the field direction. Our earlier discussion on electron swarms is valid here. On the other hand, the regions which occur near the cathode and the walls have a high field strength and resultant space charge with $n_i \gg n_e$. The electrons and ions behave here as a beam rather than as a swarm.

Ambipolar Diffusion. The radial diffusion of ions and electrons to the wall in the plasma region (4), above, does not occur at the same rate as when only one type of carrier is present. Clearly, the electrons will tend to diffuse to the walls much more rapidly than the ions leaving an excess of positive charge. A space charge field is set up which retards electrons and accelerates positive ions so that their effective diffusion rates are equalized. This process can be described in terms of the ambipolar diffusion coefficient D_a which is given approximately by:

$$D_a \doteq D_i \left[1 + \frac{T_e}{T_i} \right] \quad (6)$$

where D_i is the normal ion diffusion coefficient and T_e and T_i are the effective electron and ion temperatures, respectively.

Cathode Fall. When the current is sufficiently

small (≤ 10 mA for a cathode of area ~ 1 cm²), the discharge does not occupy the entire cathode area. The current density in the covered portion j_n is approximately constant, and the cathode fall of potential V_c is nearly independent of current and pressure. This is termed the normal cathode fall. The abnormal cathode fall occurs when $i > j_n S$, where S is the total cathode area. V_c now increases with current.

Arc and High Current Discharges. If i is increased sufficiently, a stage will eventually be reached where the cathode temperature is high enough for thermionic emission to be important. V_c now decreases with further increase in i (Fig. 7), and we are in the region of the arc discharge. The transition current clearly depends on the rate of loss of heat from the cathode and only has a definite value when the surface is uniform. In some arc discharges (e.g., Hg), the emission mechanism is probably not thermionic; these are not fully understood however.

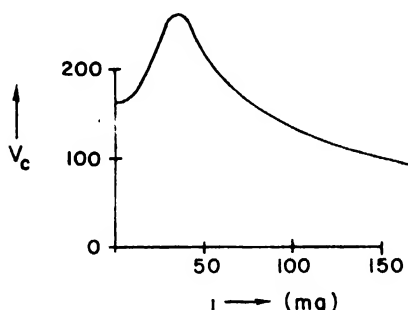


FIG. 7 Variation of cathode fall of potential with current for discharge in A at 30 torr pressure with spherical tungsten electrodes, 1.8 mm diameter.

We have assumed hitherto that the current is always sufficiently low for the magnetic field produced by the current to play an unimportant role in the discharge mechanism. At high currents this is no longer true, and the interaction of the self-magnetic field of the discharge and the current produces forces on the ionized gas comparable to the other forces acting. The required currents increase with the gas pressure p ; at normal temperatures, $i \sim 10^3$ amperes and $p \sim 1$ torr are required. The force due to the magnetic field tends to constrict the discharge, and a column so constricted is said to be pinched. This pinch effect offers a possible method of confining the hot gas to a channel remote from the walls of the containing vessel. However, a major obstacle to the achievement of a steady pinched discharge is the inherent instability of such a channel to lateral perturbations. This causes the pinched column to leave the axis of the containing tube and take up a helical path in contact with the walls. Although a suitable arrangement of magnetic fields may help to stabilize the discharge, the prospect of continuous operation of a pinched discharge does not appear promising.

J. D. SWIFT

References

- Craggs, J. D., and Meek, J. M., "Electrical Breakdown of Gases," London, Oxford University Press, 1953.
- Loeb, L. B., "Basic Processes of Gaseous Electronics," Berkeley, Cal., University of California Press, 1955.
- Jones, F. Llewellyn, "Ionisation and Breakdown in Gases," London, Methuen & Co., Ltd., 1957.
- Acton, J. R., and Swift, J. D., "Cold Cathode Discharge Tubes," London Heywood and Co., Ltd., 1963.
- von Engel, A., "Ionized Gases," London, Oxford University Press, 1955.
- Somerville, J. M., "The Electric Arc," London, Methuen & Co., Ltd., 1959.

ELECTRICAL MEASUREMENTS

In an *electrical measurement*, one is concerned with the evaluation of an electrical quantity—resistance, capacitance, inductance, current, voltage, power, energy—or of a quantity or relationship that depends on some combination of them. The measurement means may be a ratio device, such as a potentiometer or bridge in which generally similar quantities are compared; or it may be an electromechanical system in which a force is developed to produce a displacement proportional to the electrical quantity to be measured. Electrical indicating instruments, such as ammeters, voltmeters and wattmeters, use this principle. In still other measuring systems, the heating effect of an electric current is utilized.

The basis of any meaningful electrical measurement must ultimately be the *National Reference Standards* maintained by the National Bureau of Standards in Washington, D.C. The magnitudes of these National Reference Standards are assigned in terms of *absolute* measurements in which certain electrical quantities are determined in terms of appropriate mechanical quantities—the electrical system of units being related to the metric system of mechanical units in such a way that both systems have identical units of power and energy.

Two *absolute* measurements are needed to assign values to the National Reference Standards. Historically, these have been: (1) an *ohm* determination and (2) an *ampere* determination. In the ohm determination, a resistance is compared to the reactance of an inductor or capacitor at a known frequency. The magnitude of the inductor or capacitor is calculated from its measured dimensions, and the resistance value is thus assigned in terms of the mechanical units of length and time—the meter and the second. In the ampere determination, the value of a current carried by two coils is measured in terms of the force with which the coils interact. The force between the coils is opposed by the force of gravity acting on a known mass, and the ampere is assigned in terms of the mechanical units of length, mass, and time—the meter, kilogram and second. When the ampere experiment is performed, the measured current is passed through

a resistor of known value, and the potential drop is used to assign the electromotive force of a *standard cell*.

The basic National Reference Standards, in terms of which the *legal* electrical units are maintained, are groups of 1-ohm resistors and of standard cells of the cadmium sulfate type—the “Weston” saturated cell—whose values are assigned by *absolute* measurement, and whose mean is assumed constant during the interval between absolute measurements. At the National Bureau of Standards, the national resistance standard consists of ten 1-ohm resistors, fully annealed and mounted strain-free out of contact with the air in sealed containers. The national voltage standard consists of 44 Weston cells maintained at 28°C. On the basis of present evidence (1964), it is believed likely that the difference between the *legal* ohm and the defined *absolute* ohm does not greatly exceed a part in a million. The difference between the maintained *legal* volt and the defined *absolute* volt may possibly be as much as 10 to 12 parts in a million.

The *voltage divider* is the basis of many precise measurement networks. In general, it consists of a group of resistors or impedors (using resistance, capacitance, inductance, or some combination of them). Its operating principle is as follows: When the series circuit is tapped at an intermediate point but no current is drawn from the tap, the ratio of the voltage between the tap point and one terminal of the divider to the voltage impressed across the whole divider equals the ratio of the tapped resistance, (or impedance) to the total resistance (or impedance) of the divider. Modern dc potentiometers operating on this principle achieve an accuracy of a few parts in a million in comparing direct voltages and, with appropriate range-extending voltage dividers, can be used to measure voltages up to 1500 volts with an accuracy—referred to the legal volt—of 10 parts in a million. Standard cells can be intercompared to 1 to 2 parts in 10^7 , using special potentiometers designed to eliminate or minimize parasitic voltages. Direct currents can also be measured to a few parts in a million with a potentiometer, by comparing the voltage drop that the current produces in a known resistance with a known reference voltage.

In any measurement network using voltage-divider techniques, a sensitive detector such as a galvanometer must be included in the circuit element containing the tap point to indicate the absence of current in this branch. Although a number of electronic detectors have been developed for dc measurements, which are rugged and convenient to use, the D'Arsonval galvanometer, operated to make optimum use of its design characteristics, is still the most sensitive detector of voltage unbalance in potentiometers and in the usual dc bridge networks. In fact, unbalance detection at the nanovolt level in circuits having resistances up to a few hundred ohms is quite possible with the D'Arsonval galvanometer. A galvanometer consists basically of a coil of wire suspended by very fine metallic

filaments—usually flat ribbons—in a radial magnetic field. Current in the coil introduced through the suspension filaments, interacts with the magnetic field to produce a torque tending to rotate the coil. The suspension stiffness opposes the coil rotation. A light beam, reflected from a mirror fixed to the moving system, indicates the magnitude of the rotation and hence the current in the coil. The light beam can be focused on a scale for direct observation, or may be shared by two differentially connected photocells whose output is supplied to a second galvanometer, producing a greatly amplified deflection. Such a photoamplifier system may be used to achieve nanovolt response in the detector system.

Two voltage dividers may be connected in parallel to the same voltage source to form a *bridge*. Equality of divider ratios—indicated by zero voltage difference at their tap points—permits the accurate comparison of impedances and generally is relatively insensitive to minor variations in the level of the supply voltage. The best known bridge for dc resistance determination is the Wheatstone network. It incorporates two dividers, one of which provides a known ratio, the other includes the unknown resistor and a known resistor with which it is compared. The Wheatstone bridge is used for resistors, usually of value greater than 1 ohm, which have only two terminals, i.e., whose potential and current connections coincide. A second network, the Kelvin bridge, is used for resistors (usually an ohm or less) which have four terminals, i.e., whose potential and current terminals are separated. Two known ratios of identical magnitude must be provided in addition to the “divider” which incorporates the unknown and reference resistors. By employing one or the other of these bridges, and using a direct substitution method, nominally equal resistors can be compared to a few parts in a million or better in the resistance range from 10^4 to 10^{-4} ohm; at the 1-ohm level the comparison can be made to a part in 10^7 .

The more general impedance bridge is usually a 4-arm network similar to the Wheatstone bridge—or a more complicated network which, by appropriate star-delta transformations, can be reduced to an equivalent 4-arm network—in which, by proper choice of components, inductances can be intercompared or measured in terms of a combination of capacitance and resistance, or capacitances can be intercompared or determined by a combination of inductance and resistance. The accuracy of such bridges is usually limited by the stability of the reference components and the other elements of the bridge, and by how well their values are known. Additional limits may be imposed by intercoupling between bridge elements or by their coupling to nearby objects or to ground. Such coupling may consist of an ambient magnetic field inducing a parasitic emf in an inductive bridge element or even in open loops formed by connectors between bridge elements, or by capacitance or leakage resistance between elements or to a neighboring

source of potential or to ground. Such effects can be reduced or eliminated by choice and arrangement of inductive elements to avoid inductive coupling and by the use of shields maintained at appropriate potentials, to eliminate the effects of stray capacitance and leakage resistance on the bridge balance. Thus, no statement of bridge capability is possible except in the context of the individual components and their arrangement. As an extreme example of state-of-art measurements, 10-pF, 3-terminal capacitors can be compared to a part in 10^8 , in a completely shielded bridge whose ratio arms are formed of closely coupled transformer windings.

Indicating instruments are commonly used to measure current, voltage, and power, and with special circuit arrangements or transducers, they can be used to indicate other electrical quantities—resistance, frequency, phase—or nonelectrical quantities such as speed, temperature, pressure, level of illumination—in fact any quantity for which an appropriate transducer can be devised that will convert the measurand into a usable electric signal.

Direct-current instruments are usually of the permanent-magnet moving coil type, sometimes called D'Arsonval-type instruments because their operating principle is identical with that of the D'Arsonval galvanometer. A coil, suspended or pivoted in a radial field between the polepieces of a permanent magnet, tends to rotate from the mechanical interaction between its current and the field of the magnet. The turning moment of the coil is opposed by spiral springs in the case of a pivoted coil (or by the suspensions in a taut-band construction), and the equilibrium position of the coil is indicated by a pointer attached to it and moving across the instrument scale, which is marked in units of the quantity being measured. In a milliammeter, the entire current may be taken by the moving coil; in an ammeter, where the current is too large for the coil and connecting springs, a parallel circuit or shunt carries the bulk of the current, and only a fraction of it is taken by the moving coil; in a voltmeter, a large series resistance (or multiplier) is used to limit the coil current, and the indicated voltage is the product of coil current by total resistance between the instrument terminals.

In a permanent-magnet moving-coil instrument, the direction of the torque reverses with the direction of current in the moving coil. Therefore, this arrangement cannot be used for alternating-current indication. An arrangement is required for which the direction of torque is independent of current direction. A number of such arrangements are possible, and a variety of types of ac ammeters and voltmeters are used.

(1) The *rectifier* instrument makes use of four semiconductor rectifier elements arranged in a square with the input across one diagonal and a permanent-magnet moving-coil instrument connected across the other diagonal, the rectifier elements being arranged with their direction of conduction such that current in the permanent-magnet moving-coil instrument is in the same

direction for either polarity at the input terminals. This arrangement can be used either for a voltmeter or for a low-range current meter (a milliammeter), but its use for a high-range current meter (ammeter) is generally impractical. The *average* value of current or voltage over a half cycle of the alternating current is indicated. Since usually one is concerned with the effective (rms) value of current or voltage, the scale is marked in terms of the rms quantity for a sine-wave input, and wave-form errors are present for a non-sinusoidal input. Use of this type of instrument is restricted to relatively low frequencies, such as those used for transmitting power or those in the low audio-frequency range.

(2) *Thermocouple* instruments utilize the heating effect of an electric current. A fine wire or thin-walled tube is heated by the current to be measured; a thermocouple is attached to the heater element, and the temperature rise produces an emf whose value is indicated on a low-range permanent-magnet moving-coil millivoltmeter. Since the rate of heat production (and the consequent temperature rise in the heater element) is proportional to the square of the current, the indication is of *effective* (rms) value and there is no wave-form error. This type of instrument may be used for current measurement (milliammeters and ammeters) from dc to the rf range (with some constructions to 200 megahertz or more) without serious frequency errors. As a voltmeter, the use of this instrument is usually restricted to audio frequencies but, with multipliers of special design having low distributed capacitance, the useful range of thermocouple voltmeters can be extended upwards to a megahertz or more.

(3) *Electronic* instruments are used as voltmeters over a wide frequency range, extending upwards in some instances to several hundred megahertz. A variety of circuit arrangements are used. In one common arrangement, the voltage to be measured is rectified, reduced to an appropriate level by means of an attenuator, and impressed on the grid of a vacuum tube to control plate current. The magnitude of the plate current is a function of the grid voltage and is read on a dc milliammeter whose scale is marked directly in volts. In this instance, the response is to the peak value of the impressed alternating voltage, although the scale is generally marked in terms of rms volts for a sine-wave input. Thus, a substantial wave-form error may be present if the impressed voltage is not sinusoidal. In other circuit arrangements, peak, average or even rms values may be directly indicated, and it is important to know the response law of the instrument so that appropriate allowance may be made for wave-form errors if the response is not truly rms. Vacuum-tube voltmeters are an important class of instruments because of their wide frequency range and because, generally, their input impedance is quite high. In many electronic voltmeter circuits the power required for operation and indication is supplied from an auxiliary

source, and negligible power is taken from the circuit whose voltage is measured.

(4) In *moving-iron* instruments, a soft-iron piece forms the moving element. It is immersed in the field of a coil that carries the current to be measured, and its motion in the field is such as to increase the inductance of the system with increasing coil current. Since the energy stored in the magnetic field can be written $E = \frac{1}{2} I^2 L$, where L is the inductance of the system, its derivative—the torque—is $T = \partial E / \partial \theta = \frac{1}{2} I^2 \partial L / \partial \theta$, and the system's response is proportional to I^2 ; the instrument indicates rms (effective) current or voltage. By suitably shaping and arranging fixed and movable irons in the field of the coil, the instrument scale can be made very nearly linear over as much as 80 per cent of its range. Alternatively, the upper range of the scale can be greatly compressed, and a small portion of the total range can be expanded to cover much of the scale. The latter arrangement is particularly appropriate in a voltmeter that is used to monitor a voltage which is nearly constant most of the time, e.g., a line voltage.

(5) *Electrodynamic* ammeters and voltmeters have both a fixed and a moving-coil system, each of which carries the current to be measured or a fraction of it. The interaction of their fields produces a torque proportional to the square of the current if the coils are connected in series (in a voltmeter), or proportional to the product of the fixed and moving-coil currents in a parallel arrangement (in an ammeter). Thus the scale indication is of rms (effective) voltage or current as in the moving-iron instrument. Here, however, eddy-current errors are much less, and electrodynamic instruments are generally useful over an extended range of power- and low audio-frequencies. In addition, the dc response can be substantially error-free, and electrodynamic instruments are used extensively as ac-dc transfer standards to determine the ac performance of other instruments which cannot be calibrated reliably on dc.

This transfer function is an important one, since the basic standards of resistance and emf, and the potentiometer techniques for accurately measuring current and voltage, are available only on direct current. This function of electrodynamic instruments—ac-dc transfer standards—is generally confined to frequencies below a kilohertz. Over a more extended frequency range—to 20 kilohertz or more—thermocouple elements with appropriate shunts and series resistors are used as current and voltage transfer standards.

Electrodynamic instruments are also used as wattmeters to measure power at low frequencies, generally below 1 kilohertz. For this purpose, the moving coil with an appropriate series resistance is connected across the supply lines as the voltage circuit of the wattmeter. The fixed coils are connected to carry the load current. The instrument torque is proportional to the product of the currents in the fixed and moving-coil systems at any instant, and is therefore proportional to the instantaneous product of line

voltage and load current. The time integral of this product over a cycle (divided by the period) is the average power in the load, and since the moving system cannot follow variations within so short a time interval, it takes up a position that indicates average power—in effect, it performs the required integration. Thermoelement arrangements have been used to a limited extent in some laboratories to measure power through the audio-frequency range, but such instruments are not available commercially. In the rf and microwave regions, the precise measurement of power is difficult. Arrangements are available in which the power output of a circuit is absorbed in a calorimeter or bolometer, and its magnitude indicated by temperature rise, resistance change, or voltage developed across a fixed resistance.

The range of current, voltage or power that can be measured directly with indicating instruments is practically limited to a few amperes, or a few hundred volts or watts. *Instrument transformers* are used at power frequencies to extend the range of measurement capability almost indefinitely. For example, *current transformers* rated at 10 000 amperes are used in some installations, as are *voltage transformers* rated at 350 000 volts. Instrument transformers, consisting of a primary and a secondary winding coupled by a magnetic core, are designed specifically to accurately reproduce the primary current (or voltage) on a reduced scale in the secondary circuit. They are quite different from power transformers in details of design and operation, although their basic operating principle is the same. The usual *current transformer*, designed for use with a 5-ampere ammeter, or to supply the current circuit of a wattmeter or watt-hour meter, is capable of delivering only a few watts to these instrument circuits, and generally operates under nearly short-circuit conditions with its magnetic circuit at very low flux density. The usual *voltage transformer*, designed for use with a 120-volt voltmeter, or to supply the voltage circuit of a wattmeter or watt-hour meter, also generally has a relatively low power rating and operates under nearly open-circuit conditions with its magnetic circuit approaching saturation. Instrument transformers of good design can be expected to have errors of only a few hundredths or at most tenths of a per cent when operated within their design limits.

FOREST K. HARRIS

Cross-references: ALTERNATING CURRENTS, ELECTRICITY, INDUCTANCE.

ELECTRICITY

An isolated atom consists of a small nucleus, itself composed of protons and neutrons, surrounded by a cloud of electrons. The proton and the electron are the ultimate stable particles of electricity. Their charges are equal and opposite, the proton being regarded, by convention, as positive. A normal atom with its full complement of electrons is thus uncharged.

Electrostatic phenomena arise when bodies (or parts of bodies) have an excess of electrons or protons, a state usually produced by transferring electrons, e.g., by means of a battery or by rubbing two dissimilar materials together. Between two positively charged bodies (or two negatively charged ones) there is a repulsive force; between positively and negatively charged ones, an attractive force. The nature of these forces is subsumed in Coulomb's law*

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0} \frac{Q_1 Q_2}{r^2} \mathbf{r}_1 \quad (1)$$

where \mathbf{F} is the force between two charges Q_1 and Q_2 carried on bodies very small compared with their separation r (i.e., between "point" charges); \mathbf{r}_1 is a unit vector along the line joining Q_1 and Q_2 , and indicates that the force is directed along that line; and ϵ_0 is a constant that depends on the units used and takes the value $1/(36\pi \cdot 10^9)$ for the mks system of units. The forces between extended distributions of charge can be calculated by adding together (vectorially) the forces between the very small elements of charge into which the charge distributions can be imagined to be divided.

We can speak of the *electric field* of force \mathbf{E} at a point due to a system of charges and define it by means of the force \mathbf{F} on a *small* test charge q placed at the point:

$$\mathbf{E} = \mathbf{F}/q \quad (2)$$

For a point charge Q , the field at a distance r from it is then

$$\mathbf{E} = Q\mathbf{r}_1/4\pi\epsilon_0 r^2 \quad (2a)$$

If we slowly move a small charge q a small distance $d\mathbf{x}$ in a field \mathbf{E} , the electric system does work *against* the force \mathbf{F} holding the small charge in place, so the work done *on* the electrical system *by* the force is

$$dW = \mathbf{F} \cdot d\mathbf{x} = -q\mathbf{E} \cdot d\mathbf{x} \quad (3)$$

We define *electric potential difference* dV by

$$dV = dW/q = -\mathbf{E} \cdot d\mathbf{x} \quad (4)$$

The *electric flux* through a small geometrical (not necessarily material) surface of size dA is $\mathbf{E} \cdot d\mathbf{A}$ where $d\mathbf{A}$ is the vector normal to the surface. Then from equation (2a) for a surface A enclosing a charge Q , we have

$$\begin{aligned} \int_V \epsilon_0 \mathbf{E} \cdot d\mathbf{A} &= \int_Q \int_A \frac{dQ \mathbf{r}_1 \cdot d\mathbf{A}}{4\pi r^2} \\ &= \frac{1}{4\pi} \int_Q \int_\omega dQ d\omega = Q \end{aligned} \quad (5)$$

where $d\omega = \mathbf{r}_1 \cdot d\mathbf{A}/r^2$ is the solid angle subtended by dA at the element of charge dQ .

* We consider only systems which do not include dielectrics. For such see DIELECTRIC THEORY.

This is Gauss' theorem, one often of great assistance in solving electrostatic problems, for which the most difficult part can be to obtain the distribution of the charge on a system of conductors (which cannot sustain an electric field unless an emf is present or energy continuously supplied). For example, consider two parallel conducting plates of area S , a small distance t apart. Then ignoring edge effects, by symmetry \mathbf{E} is constant between the plates and perpendicular to them. Applying Gauss' theorem to a small cylinder with one end of area a in the positively charged plate (so for this end $\mathbf{E} \cdot d\mathbf{A} = 0$) and the other between the plates, we have, by Gauss' theorem, since \mathbf{E} is parallel to the sides (and so $\mathbf{E} \cdot d\mathbf{A}$ zero for the sides),

$$\int \epsilon_0 \mathbf{E} \cdot d\mathbf{A} = \epsilon_0 E a = Qa/S.$$

Then from Eq. (4) the potential difference between the plates is $V = Et = Qt/\epsilon_0 S$ and so the CAPACITANCE of the plates, defined as $C = Q/V$, is $\epsilon_0 S/t$.

The laws of electrostatics underly many phenomena such as thunderstorms and the behavior of electrons in radio and cathode-ray tubes.

Current Electricity. If two bodies at different potentials are connected by a conductor, such as a metal wire in which there are free electrons, the electrons in the wire drift under the influence of the electric field. Such movement of electrical charges gives rise to further phenomena and we speak of an *electric current*. This we can define quantitatively as the rate at which charge is passing:

$$I = dQ/dt \quad (6)$$

The current may be one of electrons only, as in a metal, in semiconductors or in electron tubes; of positive nuclei, as in an isotope-separator; or of both positive and negative charges, as in the conduction by ions (atoms that have gained or lost an electron) in liquids or in gaseous electrical discharges. (see CONDUCTIVITY, ELECTRICAL; ELECTRIC DISCHARGES IN GASES). Note that *electrons* flowing in the *positive* direction give rise, by our convention of signs, to a *negative* current and that in a metal, semiconductor, or conducting liquid, the velocity at which the electrons or ions drift is quite slow, less than 1 cm/sec even for current densities in a metal as high as 10^4 amperes cm^{-2} . In vacuum devices, such as cathode-ray tubes, the speed of the electrons approaches that of light.

If a wire joins two electrostatic charges, the current lasts for a short time only, but it may be maintained by means of some source of energy, such as a battery, a generator, a thermocouple, or a solar photoelectric cell.

When a current I flows under the influence of a potential difference V , the moving charges—electrons in metals, ions in solutions—are impeded by collision with the atoms in the conducting metal or liquid. The charges give up to the atoms the energy they acquired as they moved in the electric field, and electrical power is converted into other forms, — for instance, into heat in the case

of a metal wire. From Eqs. (4) and (6), this power P is given by

$$P = \frac{dW}{dt} = \frac{d(VQ)}{dt} = V \frac{dQ}{dt} = VI \quad (7)$$

A metal wire at constant temperature (but not all conductors) obeys Ohm's law: that is, $V = RI$, where R is a constant for the wire known as its resistance (in ohms if V and I are in volts and amperes). In this case

$$P = RI^2 = V^2/R \quad (8)$$

and this is the rate at which heat is generated. The heat in most forms of electric heaters (including filament lamps) is generated in this way. (For an explanation of how current flows in more complicated circuits see "CIRCUITRY.")

Accompanying a current I there is always a magnetic field \mathbf{H} which shows itself by the force exerted between the current-carrying conductor and a magnet or another electric current. At a point P a distance a from a long, thin, straight wire carrying current I , the field \mathbf{H} is perpendicular to both the wire and the radius from the wire to P , so that the "lines of force" are circles concentric with the wire and perpendicular to it. The field is of magnitude

$$H = I/2\pi a \quad (9)$$

For a long cylindrical coil, $H = nI$, where n is the number of turns per unit length. In general, we have the relation that

$$\oint \mathbf{H} \cdot d\mathbf{s} = NI \quad (10)$$

where N is the total number of turns, each carrying current I , enclosed by the path s . This holds true whether or not the path traverses magnetic materials. Associated with the field \mathbf{H} , there is a magnetic flux density \mathbf{B} ,

$$\mathbf{B} = \mu_0 \mu_r \mathbf{H} \quad (11)$$

where μ_0 is a constant of value $4\pi \times 10^{-7}$ in mks units and μ_r is the relative permeability. μ_r is not necessarily a constant, but it has the value 1 for free space and is of the order of 1000 for magnetic materials.

The magnetic flux density satisfies Gauss' theorem [see Eq. (5)], but because magnetic poles cannot be isolated in the way electric charges can be, the theorem takes the form

$$\int_A \mathbf{B} \cdot d\mathbf{A} = 0 \quad (12)$$

where the integral is taken over the geometrical surface A . Equations (10), (11) and (12) are sufficient, in principle, for the calculation of the magnetic field of any steady current.

The force between a current I , traversing a short length $d\mathbf{l}$ in a magnetic field of flux density \mathbf{B} , and the field is

$$I d\mathbf{l} \times \mathbf{B} \quad (13)$$

so for a wire of length l in a field perpendicular to it, the force is lB perpendicular to both wire and field. The force between two long parallel wires in free space, each carrying a current I and distance a apart, is, from Eqs. (9) and (13),

$$F = 2 \times 10^{-7} I^2 l / a \quad \text{newtons/m} \quad (14)$$

For currents in the same direction it is an attraction. This equation defines the mks unit of current though that unit is *established* by means of the larger force between coils carrying a common current I , for which $F = KI^2$, where K is a factor that can be calculated accurately from the geometry of the coils. The force between a current-carrying coil and magnetic material, or another current, is the basis of many electrical instruments such as ammeters, of electric motors and of other electromechanical devices.

If we think of the current as consisting of electrons, or ions, of charge Q , moving with velocity \mathbf{v} along a wire or in a vacuum, these experience a force in a magnetic field \mathbf{B} given by

$$\mathbf{F} = Q\mathbf{v} \times \mathbf{B} \quad (15)$$

a relation that is required for the explanation of the behavior of electrons in a magnetic electron microscope, of ions in a mass spectrometer, and of cosmic rays in the earth's magnetic field. Further, if a wire, with its electrons free except in so far as they are confined to the wire, and a magnetic field \mathbf{B} are caused to move relative to one another with velocity \mathbf{v} , by virtue of Eq. (15) the electrons will suffer a force *along* the wire; an INDUCED ELECTROMOTIVE FORCE is set up. If the ends of the wire are connected by a conductor, a current will flow. This is the phenomenon of *electromagnetic induction* and the basis of equipment such as transformers and electric generators.

When an electric current changes rapidly, electromagnetic waves are radiated. If the current is an alternating one, the power radiated, for a current of given size, increases as the square of the frequency. Such waves, at frequencies of 100 kc/sec or more, are the basis of radio communication.

Though electrostatic phenomena are important—in radio tubes, for instance—current electricity is vital in the modern world. Small currents can be maintained by batteries and secondary cells, but the very large amounts of power used in a modern industrial state must be obtained mainly by electromagnetic machinery (see ELECTRIC POWER GENERATION), usually as ALTERNATING CURRENTS.

E. S. SHIRE

References

- Kip, A. H., "Fundamentals of Electricity and Magnetism," New York, McGraw-Hill Book Co., 1962.
- Shire, E. S., "Classical Electricity and Magnetism," Cambridge, Cambridge University Press, 1960.
- Panofsky, W. K. H. and Phillips, M., "Classical Electricity and Magnetism," 2nd Ed., 1962 Addison-Wesley, Reading, Mass.

Cross-references: CAPACITANCE; CIRCUITRY; CONDUCTIVITY, ELECTRICAL; DIELECTRIC THEORY; ELECTRIC POWER GENERATION; ELECTRICAL DISCHARGES IN GASES; INDUCED ELECTROMOTIVE FORCE; INDUCTANCE; POTENTIAL; STATIC ELECTRICITY.

ELECTROACOUSTICS

Introduction. Electroacoustics is concerned with the principles (*transduction processes*) and devices (*transducers*) by which electrical energy may be converted into acoustic energy and vice versa. Consider the familiar electrodynamic transducer. A periodic electric current passing through a coil interacts with a steady radial magnetic flux causing the coil to vibrate. The coil in turn drives a diaphragm which radiates sound waves from one side. (The other side is usually enclosed to avoid cancellation of the acoustic output.) The entire process is *reversible* since sound waves striking the diaphragm set up a periodic variation in air pressure adjacent to the diaphragm causing it to vibrate. As the moving coil cuts the magnetic flux, an emf is generated which causes a current to flow when a load is connected to the coil terminals.

Many, but not all, types of transducer are similarly reversible. A reversible transducer may be made to perform sending and receiving functions successively in such a manner that an absolute sensitivity may be determined (*reciprocity* calibration).

The electrodynamic transducer may further be classified as *passive* since all of the energy appearing in the acoustic load is derived from the electrical input energy, and *linear* in the sense that there is a substantially linear relationship between the input and output variables (electric current and acoustic pressure in the present case).

Transduction Processes. *Irreversible Transducers.* These depend on a variety of special effects of which the best known is (a) the variation of surface contact electrical resistance with pressure (carbon microphone). Other effects are (b) the variation of bulk resistance with elastic strain (piezoresistance), (c) variation of transistor parameters with strain, (d) cooling effect of periodic air movement (hot wire microphone), (e) pressure wave generated by an electrical spark, (f) dependence of air pressure on level of corona discharge (ionophone).

Reversible Transducers. A important class of reversible transducer depends on relative movement of suitable components linked by an electric or magnetic field traversing a gap. Examples are (a) the electrodynamic transducer already described; (b) electrostatic depending on the relative movement of charged condenser plates; (c) magnetic or variable reluctance depending on relative movement of magnetic poles in a magnetic circuit linked with a fixed coil.

Other *reversible transducers* are dependent on dimensional changes connected with the state of magnetic or electric polarization of certain crystalline materials (piezomagnetism and piezoelectricity). Since strain may be longitudinal or

shear and since both strain and polarization are directional quantities, many possible relationships between strain and polarization exist. The behavior of an X-cut quartz disk may serve as an illustration. When such a disk is axially compressed, electric charges appear on the plane surfaces. Conversely, if a potential difference is established between the two surfaces, contraction or expansion occurs depending on the direction of the electric field. Other important single-crystal piezoelectric materials are ammonium dihydrogen phosphate (ADP) and Rochelle salt. During the past decade, polycrystalline ceramic materials based on barium titanate and lead zirconate titanate have replaced single-crystal materials in many applications. These materials are ferroelectric and, when prepolarized, exhibit piezoelectric behavior.

To date, only polycrystalline piezomagnetic materials (often termed magnetostrictive) have been found useful. Some are metals such as nickel and permendur. Others are ferrite ceramics [basic composition: $(\text{NiO})(\text{Fe}_2\text{O}_3)$] which have such a high electrical resistivity that eddy current losses are negligible making lamination unnecessary.

Electromechanical Coupling. Transducer performance is closely connected with the tightness of coupling between mechanical and electrical aspects. Consider a piezoelectric disk which is compressed by putting in *mechanical* energy W_m . The appearance of surface charges shows that *electrical* energy W_e is stored in the self capacitance and is available when an external circuit is connected to suitable electrodes. The ratio W_e/W_m (electromechanical coupling coefficient) sets a limit to the efficiency for a given bandwidth (frequency range). The coefficient may reach 70 per cent for lead zirconate titanate.

Transducer Design. Impedance matching is of primary importance in electroacoustics. It may be likened to the choice of gear ratio and wheel size in automobile design. Impedance matching is generally closely related to transducer parameters such as beam width of projected or received sound and frequency response, as well as efficiency. The many available matching techniques include (a) resonance, (b) horn systems (acoustic transformers), (c) lever systems (mechanical transformers). In the direct radiator electrodynamic loudspeaker, the diaphragm is made large enough to interact with the acoustic medium (air) and yet small enough in relation to the sound wavelength (at low frequencies, at least) to ensure uniform projection of sound over a wide angle. In the condenser loudspeaker, a large transducer area compensates for the weakness of electrostatic forces. In the underwater sonar projector, slabs of piezoelectric ceramic may be sandwiched between metal plates to form a resonant device which radiates a narrow beam of sound with high efficiency over a narrow frequency range.

Recent Advances. Discoveries in the solid-state field have already provided new piezoelectric

and piezomagnetic materials. Recently the interaction between free electrons and phonons in cadmium sulfide has provided a means of amplifying ultrasonic waves. The polarization voltage hitherto required in condenser transducers may now be replaced by a prepolarized plastic film which stores the required charge indefinitely. Finally, intense acoustic waves at frequencies up to 5000 Mc/sec have been generated by electrostrictive processes accompanying the passage of laser beams through liquids and solids.

E. A. G. SHAW

References

- Hunt, F. V., "Electro-acoustics," Cambridge, Harvard University Press and New York, John Wiley & Sons, Inc., 1954.
 Beranek, L. L., "Acoustics," New York, McGraw-Hill Book Co., Inc., 1954.
 Bradfield, G., "Ultrasonic Electroacoustics," *Proceedings of the First International Congress on Acoustics*, 171 (1953).
 Hutson, A. R., McFee, J. H., and White, D. L., "Ultrasonic Amplification in CdS," *Phys. Rev. Letters*, 7, 237 (1961).
 Sessler, G. M., and West, J. E., "Self-biased Condenser Microphone with High Capacitance," *J. Acoust. Soc. Am.*, 34, 1774 (1962).
 Garmire, E., and Townes, C. H., "Stimulated Brillouin Scattering in Liquids," *Applied Physics Letters*, 5, 84 (1964).

Cross-references: ACOUSTICS, NOISE, RESONANCE, ULTRASONICS, LASER.

ELECTROCHEMISTRY

Electrochemistry is that branch of science which deals with the interconversion of chemical and electrical energies, i.e., with chemical changes produced by electricity as in electrolysis or with the production of electricity by chemical action as in electric cells or batteries. The science of electrochemistry began about the turn of the eighteenth century. In 1796 Alessandro Volta observed that an electric current was produced if unlike metals separated by paper or hide moistened with water or a salt solution were brought into contact. Volta used the sensation of pain to detect the electric current. His observation was similar to that observed ten years earlier by Luigi Galvani who noted that a frog's leg could be made to twitch if copper and iron, attached respectively to a nerve and a muscle, were brought into contact.

In his original design Volta stacked couples of unlike metals one upon another in order to increase the intensity of the current. This arrangement became known as the "voltaic pile." He studied many metallic combinations and was able to arrange the metals in an "electromotive series" in which each metal was positive when connected to the one below it in the series.

Volta's pile was the precursor of modern batteries (see BATTERIES).

In 1800 William Nicholson and Anthony Carlisle decomposed water into hydrogen and oxygen by an electric current supplied by a voltaic pile. Whereas Volta had produced electricity from chemical action these experimenters reversed the process and utilized electricity to produce chemical changes. In 1807 Sir Humphry Davy discovered two new elements, potassium and sodium, by the electrolysis of the respective solid hydroxides, utilizing a voltaic pile as the source of electric power. These electrolytic processes were the forerunners of the many industrial electrolytic processes used today to obtain aluminum, chlorine, hydrogen, or oxygen, for example, or in the electroplating of metals such as silver or chromium.

Since in the interconversion of electrical and chemical energies, electrical energy flows to or from the system in which chemical changes take place, it is essential that the system be, in large part, conducting or consist of electrical conductors. These are of two general types—electronic and electrolytic—though some materials exhibit both types of conduction. Metals are the most common electronic conductors. Typical electrolytic conductors are molten salts and solutions of acids, bases, and salts.

A current of electricity in an electronic conductor is due to a stream of electrons, particles of subatomic size, and the current causes no net transfer of matter. The flow is, therefore, in a direction contrary to what is conventionally known as the "direction of the current." In electrolytic conductors, the carriers are charged particles of atomic or molecular size called *ions*, and under a potential gradient, a transfer of matter occurs.

An electrolytic solution contains an equivalent quantity of positively and negatively charged ions whereby electroneutrality prevails. Under a potential gradient, the positive and negative ions move in opposite directions with their own characteristic velocities and each accordingly carries a different fraction of the total current through any one solution. Each fraction is referred to as the ionic transference number. Furthermore, the velocity increases with temperature causing a corresponding increase in electrolytic conductivity. This characteristic is opposite to that observed for most electronic conductors which show less conductivity as their temperature is increased.

The concept that charged particles are responsible for the transport of electric charges through electrolytic solutions was accepted early in the history of electrochemistry. The existence of ions was first postulated by Michael Faraday in 1834; he called negative ions "anions" and positive ones "cations." In 1853, Hittorf showed that ions move with different velocities and exist as separate entities and not momentarily as believed by Faraday. In 1887, Svante Arrhenius postulated that solute molecules dissociated spontaneously into *free ions* having no influence on each other. However, it is known that ions

are subject to coulombic forces, and only at infinite dilution do ions behave ideally, i.e., independently of other ions in the solution. Ionization is influenced by the nature of the solvent and solute, the ion size, and solute-solvent interaction. The dielectric constant and viscosity of the solvent play dominant roles in conductivity. The higher the dielectric constant, the less are the electrostatic forces between ions and the greater is the conductivity. The higher the viscosity of the solvent, the greater are the frictional forces between ions and solvent molecules and the lower is the electrolytic conductivity.

In 1923 Debye and Hückel presented a theory which took into account the effect of coulombic forces between ions. They introduced the concept of the ion atmosphere, in which at some radial distance r from a central ion, there is, on a time average, an ionic cloud of opposite charge which sets up a potential field whose magnitude depends on the magnitude of r . This interionic attraction leads to two effects on the electrolytic conductivity. Under a potential gradient, an ion moves in a certain direction. However, the ion cloud, being of opposite sign will tend to move in the opposite direction, and because of its attraction for the central ion, will have a retarding effect on the ion velocity and thereby lead to a lowering in the electrolytic conductivity. On the other hand, the central ion will tend to pull the ion cloud with it to a new location. The ion atmosphere will adjust to its new location in time, but not instantaneously, and the delay results in a dissymmetry in the potential field around the ion. This also causes a lowering in the conductance of the solution. These effects become more pronounced as the concentration of the solution is increased; for dilute solutions, below about 0.1 molal, the equivalent conductance decreases with the square root of the concentration. For more concentrated solutions, the relation between conductivity and concentration is much more complex and depends more specifically on individual solute properties.

Interionic attraction in dilute solutions also leads to an effective ionic concentration or activity which is less than the stoichiometric value. The *activity* of an ion species is its thermodynamic concentration, i.e., the ion concentration corrected for the deviation from ideal behavior. For dilute solutions the activity of ions is less than one, for concentrated solutions it may be greater than one. It is the ionic activity that is used in expressing the variation of electrode potentials, and other electrochemical phenomena, with composition.

When electricity passes through a circuit consisting of both types of electrical conductors, a chemical reaction always occurs at their interface. These reactions are electrochemical. When electrons flow from the electrolytic conductor, oxidation occurs at the interface while reduction occurs if electrons flow in the opposite direction. These electronic-electrolytic interfaces are referred to as *electrodes*; those at which oxidation occurs are known as *anodes* and those at which

reduction occurs, as *cathodes*. An anode is also defined as that electrode by which "conventional" current enters an electrolytic solution, a cathode as that electrode by which "conventional" current leaves. Positive ions, for example, ions of hydrogen and the metals, are called *cations* while negative ions, for example, acid radicals and ions of nonmetals are called *anions*.

In 1833, Michael Faraday enunciated two laws of electrolysis which give the relation between chemical changes and the product of the current and time, i.e., the total charge (coulombs) passed through a solution. These laws are: (1) the amount of chemical change, e.g., chemical decomposition, dissolution, deposition, oxidation, or reduction, produced by an electric current is directly proportional to the quantity of electricity passed through the solution; (2) the amounts of different substances decomposed, dissolved, deposited, oxidized, or reduced are proportional to their chemical equivalent weights. A chemical equivalent weight of an element or a radical is given by the atomic or molecular weight of the element or radical divided by its valence; the valence used depends on the electrochemical reaction involved. The electric charge on an ion is equal to the electronic charge or some integral multiple of it. Accordingly, a univalent negative ion has a charge equal in magnitude and of the same sign as a single electron, and its chemical equivalent weight is equal to its atomic weight, if an element, or to its molecular weight, if a radical. A trivalent ion has $+3$ or -3 electronic charges, depending on whether it is a positive or a negative trivalent ion. For trivalent ions, then, the equivalent weight would be equal to its atomic weight, if an element, or to its molecular weight, if a radical, divided by three.

The quantity of electricity required to produce a gram-equivalent weight of chemical change is known as the *faraday*. A faraday corresponds, then, to an *Avogadro number of charges*. The most accurate determination of the faraday has been made by a silver-perchloric acid coulometer in which the amount of silver electrolytically dissolved in an aqueous solution of perchloric acid is measured. This method gives 96487 coulombs (or ampere-seconds) per gram-equivalent for the faraday or the unified C^{12} scale of atomic weights adopted in 1961 by the International Commission on Atomic Weights.

The *electrochemical equivalent* or, preferably, the *coulomb equivalent* of an element or radical is that weight in grams which is equivalent to one coulomb of electricity and is given by the gram-equivalent weight divided by the faraday (96487 coulombs per gram-equivalent); for example, the electrochemical equivalent of silver is given by $107.870/96487$ or 0.00111797 g/coulomb where 107.870 is the atomic weight of silver based on the unified C^{12} scale adopted in 1961. The electrochemical equivalents of other elements may be calculated in like fashion.

In electrolysis and in any electric cell or battery, there is an electromotive force (emf) or voltage across the terminals. This emf is expressed

in the practical unit, the volt, which is equal to the electromagnetic unit in the meter-kilogram-second system. In any one cell, the emf is the sum of the potentials of the two electrodes and of any liquid-junction potentials that may be present. Neither of the individual electrode potentials can be evaluated without reference to a chosen reference electrode of assigned value. For this purpose, the hydrogen electrode has been universally adopted and is arbitrarily assigned a zero potential for all temperatures when the hydrogen ion is at unit activity and the hydrogen gas is at atmospheric pressure. A hydrogen electrode consists of a stream of hydrogen gas bubbling over platinized platinum or gold foil and immersed in a solution containing hydrogen ions; the electrochemical reaction is: $1/2\text{H}_2(\text{gas}) \rightleftharpoons \text{H}^+(\text{solution}) + e$, where e represents the electron. The potential of the hydrogen electrode, E_{H} , as a function of hydrogen ion concentration and hydrogen-gas pressure is given by

$$E_{\text{H}} - E_{\text{H}}^0 = (RT/nF) \ln (a_{\text{H}^+} / p_{\text{H}_2}^{1/2}) \\ = E_{\text{H}}^0 - (RT/nF) \ln (c_{\text{H}^+} f_{\text{H}^+} / p_{\text{H}_2}^{1/2}),$$

where E_{H}^0 is the standard quantity assigned a value of zero, R is the gas constant, T the absolute temperature, n the number of equivalents, F the faraday, p_{H_2} the pressure of hydrogen, and a_{H^+} , c_{H^+} , and f_{H^+} , respectively, the activity, concentration, and activity coefficient of hydrogen ions. When a_{H^+} and $p_{\text{H}_2}^{1/2}$ equal one, $E_{\text{H}} = E_{\text{H}}^0$. For very dilute solutions below 0.01 molal f_{H^+} may be taken as unity without appreciable error.

The standard potentials, E^0 , of other electrodes are obtained by direct or indirect comparison with the hydrogen electrode. Values thus obtained at 25°C for some typical elements are listed in Table I.

The reducing power of the elements decreases on going down the column. These values are for the ions at unit activity, and reversible or thermodynamic values as a function of metal or radical concentration are given by equations similar to the one above. For the general reaction: $\text{M} \rightleftharpoons \text{M}^{n+} + ne$, the potential is given by $E_{\text{M}} = E_{\text{M}}^0 - (RT/nF) \ln a_{\text{M}^{n+}}$.

In electrolysis, at very low current densities, the potentials of the electrodes approximate in magnitude their reversible values and deviate somewhat from these values because of an IR drop in the solution and possible concentration polarization (the concentration at the electrode surface may differ from that in the bulk of the solution). Also for high current densities, especially for the generation of gases such as hydrogen, oxygen or chlorine, the voltage required exceeds the reversible voltage; the excess voltage is known as overvoltage, or overpotential for a single electrode, and arises from energy barriers at the electrode. Overpotential, in general, increases logarithmically with an increase in current density.

In addition to the above topics, it is frequently customary to include under electrochemistry: (1) processes for which the net reaction is physical transfer, e.g., concentration cells; (2) electrokinetic phenomena, e.g., electrophoresis, electroosmosis, streaming potential; (3) properties of electrolytic solutions if determined by electrochemical or other means, e.g., activity coefficients and hydrogen ion concentration; (4) processes in which electrical energy is first converted to heat which in turn causes a chemical reaction to occur that would not do so spontaneously at ordinary temperature. The first three are frequently considered a portion of physical chemistry, and the last one is a part of electrothermics or electrometallurgy.

The passage of electricity through gases is sometimes included under electrochemistry. However, in electrical discharges in gases, the principles are entirely different from what they are in the electrolysis of electrolytic solutions. Whereas in the latter, ionic dissociation occurs spontaneously as a result of forces between solvent and solute and without the application of an external field, for gases relatively high voltages must be applied to accelerate the electrons from the electrode to a velocity at which they can ionize the gas molecules they strike. In this case, the resulting chemical reaction taking place between ions, free radicals, and molecules occurs in the gas phase and not at the electrodes as in

TABLE I. SOME STANDARD ELECTRODE POTENTIALS AT 25°C

Electrode	Potential (V)	Electrode	Potential (V)
$\text{Li} \rightleftharpoons \text{Li}^+ + e$	-3.045	$\text{Cu} \rightleftharpoons \text{Cu}^{++} + 2e$	+0.337
$\text{Ca} \rightleftharpoons \text{Ca}^{++} + 2e$	-2.87	$\text{Cu} \rightleftharpoons \text{Cu}^+ + e$	+0.521
$\text{Na} \rightleftharpoons \text{Na}^+ + e$	-2.714	$2\text{I}^- \rightleftharpoons \text{I}_2 + 2e$	+0.536
$\text{Mg} \rightleftharpoons \text{Mg}^{++} + 2e$	-2.37	$2\text{Hg} \rightleftharpoons \text{Hg}_2^{++} + 2e$	+0.789
$\text{Al} \rightleftharpoons \text{Al}^{+++} + 3e$	-1.66	$\text{Ag} \rightleftharpoons \text{Ag}^+ + e$	+0.799
$\text{Mn} \rightleftharpoons \text{Mn}^{++} + 2e$	-1.18	$\text{Pd} \rightleftharpoons \text{Pd}^{++} + 2e$	+0.987
$\text{Zn} \rightleftharpoons \text{Zn}^{++} + 2e$	-0.763	$\text{Pt} \rightleftharpoons \text{Pt}^{++} + 2e$	+1.20
$\text{Fe} \rightleftharpoons \text{Fe}^{++} + 2e$	-0.440	$2\text{Cl}^- \rightleftharpoons \text{Cl}_2 + 2e$	+1.36
$\text{Ni} \rightleftharpoons \text{Ni}^{++} + 2e$	-0.250	$\text{Au} \rightleftharpoons \text{Au}^+ + e$	+1.68
$\text{H}_2 \rightleftharpoons 2\text{H}^+ + 2e$	0.000	$2\text{F}^- \rightleftharpoons \text{F}_2 + 2e$	+2.87

the electrolysis of solutions. Studies of the electrical conduction of gases, accordingly, are generally considered under the physics of gases.

Electrochemistry finds wide application. In addition to industrial electrolytic processes, electroplating, and the manufacture and use of batteries already mentioned, the principles of electrochemistry are used in chemical analysis, e.g., polarography, and electrometric or conductometric titrations; in chemical synthesis, e.g., dyestuffs, fertilizers, plastics, insecticides; in biology and medicine, e.g., electrophoretic separation of proteins, membrane potentials; in metallurgy, e.g., corrosion prevention, electrorefining; and in electricity, e.g., electrolytic rectifiers, electrolytic capacitors.

WALTER J. HAMER

References

- Pitzer, K. S., and Brewer, Leo in Lewis, G. N., and Randall, M., Eds., "Thermodynamics," second edition, Ch. 24, New York, McGraw-Hill Book Company, Inc., 1961.
- Delahay, Paul, "New Instrumental Methods in Electrochemistry," New York, Interscience Publishers, Inc., 1954.
- Hamer, W. J., Ed., "The Structure of Electrolytic Solutions," New York, John Wiley & Sons, Inc., 1959.

Cross-references: BATTERIES; CONDUCTIVITY, ELECTRICAL; DIELECTRIC THEORY; ELECTRICAL DISCHARGES IN GASES; IONIZATION; MOLECULAR WEIGHT; POTENTIAL; VISCOSITY.

ELECTROLUMINESCENCE

Electroluminescence is luminescence excited by electric fields or currents in the absence of bombardment or other means of excitation. Several different types of electroluminescence can be distinguished.

The first observation of what is known as "recombination or injection electroluminescence" was made in 1923 by Lossev, who found that when point electrodes were placed on certain silicon carbide crystals and current passed through them, light was often emitted. Explanation of this emission has been possible only with the development of modern semiconductor theory. If minority charge carriers are injected into a semiconductor, i.e., electrons are injected into *p*-type material or "positive holes" into *n*-type material, they recombine spontaneously with the majority carriers existing in the material. If some of these recombinations result in the emission of radiation, electroluminescence results. Minority-carrier injection may occur not only at point contacts but also at broad area rectifying junctions; in this case the junction must be biased in the forward or "easy-flow" direction, and the electric field in the junction is lower when the voltage is applied than in its absence.

The recombinations in this type of electroluminescence may occur directly between energy

bands of the host material ("recombination radiation" or "edge emission") or by means of impurity or activator centers. In general, the emission intensity is a linear function of the injected current; the emission is "current controlled." This type of emission has now been observed in a wide variety of materials, including SiC, diamond, Si, Ge, CdS, ZnS, ZnSe, ZnO, and many of the so-called III-V compounds such as AlN, GaSb, GaAs, GaP, InP, and InSb. The emission of many of these materials lies in the infrared region of the spectrum. In some cases the efficiency is also very low, values of one emitted quantum per million injected carriers often being observed. However, in some materials the efficiency is high and approaches one emitted photon for each carrier passing across the junction. In GaAs, InAs, InP, and InSb, sufficient radiation intensity may be obtained to cause stimulated rather than normal (spontaneous) emission, resulting in an electrically excited, solid-state LASER.

Another kind of electroluminescence was first observed in 1936 by Destriau. He observed the emission of light from a specially prepared zinc sulfide phosphor suspended in an insulating oil and subjected to an intense alternating electric field by means of capacitor-like electrodes; the emission is "field controlled." Today the phosphor powder is usually embedded in a solid organic (plastic) or inorganic (glass) medium. Thin films of tin oxide or of metal are used to provide transparent, electrically conductive electrodes. Such "cells" can be made with a base of glass, metal, or flexible or rigid plastic material.

This type of electroluminescence, which has been called the "Destriau effect" or "acceleration-collision electroluminescence" has also been observed in ZnSe, Zn₂SiO₄:Mn, BaTiO₃, TiO₂, BN, and some organic materials. The best electroluminescent phosphors, however, are still ZnS and related materials; copper is the most common activator, although ZnS:Cu,Mn is also used. The emission or recombination process in these materials is the same as for excitation by ultraviolet radiation or by cathode rays. The excitation process, however, is quite different. It has been found that to prepare good electroluminescent phosphors, it is necessary to make the material very nonuniform electrically by introducing, on the surface or in defects in the interior of the ZnS particles, segregations of a relatively good conductor, such as ZnO or Cu₂S. The local electric field strength in the neighborhood of these segregations may be a hundred or more times the applied field strength, which is already of the order of 10⁴ to 10⁵ volts/cm. It has been commonly assumed that under the action of these intense local fields, electrons are liberated and accelerated to acquire considerable energy from the field; some of these energetic electrons may collide with and excite or ionize the activator or luminescence centers. However, the mechanism of electroluminescence in ZnS is still not completely understood. Minority carrier injection may also be important in this material.

If a p - n junction is biased in the reverse direction, which results in high internal field strength, other types of emission can occur. For example, the presence of energetic ("hot") carriers can result in emission at energies greater than the forbidden band gap of the material; this has been called "avalanche emission." In this way visible radiation can be generated in germanium or silicon. Emission at energies less than the band gap can also occur. This is attributed to intraband transitions; it is therefore "deceleration radiation" resulting from a change in kinetic energy of the carriers. The efficiency in this case is expected to be quite low.

If metal electrodes such as Al or Ta are immersed in suitable electrolytes and current is passed between them, light emission may also be observed. This "galvanoluminescence" is often electroluminescence in a thin oxide layer formed on the electrode by electrolytic action.

Electroluminescence in ZnS. The intensity of electroluminescence in ZnS increases rapidly as the applied alternating voltage is increased, and also increases slightly less than linearly as the frequency is increased until a region of saturation is approached at frequencies of the order of 100,000 cps. Very high brightness may be achieved (2500 foot-lamberts or more at 20,000 cps). Maximum efficiency is not achieved at the same conditions as maximum brightness. The efficiency of electroluminescence (2.5 to 3.0 per cent) is still low compared to other light sources. Some comparative figures are: electroluminescence, 10 lumens/watt; incandescence, 16 lumens/watt; fluorescent lamp, 70 lumens/watt. Electroluminescence, however, is an area source of light, in contrast to an incandescent lamp, which is essentially a point source, or a fluorescent lamp, which is a line source. Electroluminescent lamps may be made only a fraction of an inch in thickness and of any size or shape. Some electroluminescent phosphors also exhibit a change in emission color as the frequency of the applied voltage is varied.

Another feature of electroluminescent ZnS phosphors of practical importance is the decrease in output during operation. This deterioration is strongly influenced by humidity, temperature, and operating frequency. High temperature and high frequency greatly accelerate the deterioration; the time required to produce a given loss is usually inversely proportional to the frequency. Moisture must be carefully excluded. Lamps have been made with an initial output of 100 foot-lamberts and a life to 50 per cent of this value of 1000 hours (operation at 1000 to 3000 cps). The time required for the output to "build up" after application of a sinusoidal voltage is of the order of a few cycles and hence decreases as the frequency is increased; this time may be decreased by application of a dc bias. The decay time after excitation is independent of frequency.

Phosphors with controlled brightness-voltage characteristics may be prepared for specialized electronic applications. In addition to powder phosphors, continuous films of ZnS a few microns

thick may be prepared. These films operate on alternating as well as direct current and at comparatively low voltage (of the order of 20 volts). Superposition of alternating and direct current can lead to interesting interaction and control of output by small signals.

In electronic applications, electroluminescent phosphors ("electroluminors") may be advantageously used in conjunction with other solid-state components such as nonlinear resistors, photoconductors, magnetic devices, and ferro-electrics for voltage distribution and control. Uses range from single-element control circuits to complicated logic circuits, multielement storage devices, shift registers, light and x-ray image amplifiers and storage panels, and thin large-area image display devices for radar or television. Specially segmented lamps for display of numerical, alphabetical, or other types of information are also available.

Electrophotoluminescence. In addition to electroluminescence proper, other interesting effects occur when electric fields are applied to a phosphor which is concurrently, or has been previously, excited by other means. Such effects are usually termed electrophotoluminescence. One such phenomenon is the "Gudden and Pohl effect," discovered in 1920. Here the phosphor is excited (by ultraviolet radiation, for example) and then an electric field is applied during the afterglow or phosphorescence, or even after the emission has decayed below the limit of detection. A burst of emission is observed with some materials, even those which do not exhibit electroluminescence. In this case, the field acts on electrons in traps that are responsible for the phosphorescence.

Most phosphors, if continuously excited by ultraviolet radiation, x-rays, or cathode rays, show a decrease in emission if an electric field is simultaneously applied. This "field quenching" was first observed by Déchéne in 1935. In 1954, Destriau discovered that some ZnCdS:Mn phosphors excited by x-rays show the opposite effect; i.e., their emission is increased by the application of an electric field. Since that time, a similar enhancement effect has been observed for excitation by light, ultraviolet radiation, cathode-rays, and α -particles. These effects may find application in radiation converters, light amplifiers, and particle detectors.

HENRY F. IVEY

References

- Henisch, H. K., "Electroluminescence," New York, Pergamon Press, 1962.
- Ivey, H. F., "Electroluminescence and Related Effects," New York, Academic Press, 1963.
- Lax, B., "Semiconductor Lasers," *Science*, **141**, 1247-1255 (1963).
- Henisch, H. K., "Electroluminescence". *Rept. Prog. Phys.*, **27**, 364-405 (1964).

Cross-references: LASERS, LUMINESCENCE, SEMICONDUCTORS.

ELECTROMAGNETIC THEORY

The task of electromagnetic theory is to account for the effects of electrical charges in various states of motion. Although historically electromagnetic theory was developed from Coulomb's celebrated law, it is at present more economic to develop it differently^{9,10}. The macroscopic effects are described with remarkable accuracy by the following set of equations (rationalized mks system of units)

$$\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B} \quad (1)$$

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \quad (2)$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \quad (3)$$

$$\nabla \cdot \mathbf{E} = \frac{\partial \mathbf{B}}{\partial t} \quad (4)$$

$$\mathbf{D} = \epsilon_0 \mathbf{E} \quad (5)$$

$$\mathbf{B} = \mu_0 \mathbf{H} \quad (6)$$

$$\mathbf{J} = \sigma \mathbf{E} \quad (7)$$

provided the functional relationships indicated in Eqs. (5), (6), and (7) are known explicitly. With these equations and the laws of mechanics, classical electromagnetic theory becomes essentially a branch of applied mathematics.

Equation (1), sometimes known as the Lorentz force equation, defines the field quantities, \mathbf{E} , the electric field intensity, and \mathbf{B} , the magnetic induction, in terms of an observable, the force \mathbf{F} on a charge q . In Eq. (1), \mathbf{v} is the velocity of the charge relative to the observer. Equation (2) is a statement of the law of conservation of electric charge in terms of the charge density ρ and the total current density \mathbf{J} . Equation (3) is the differential form of Ampère's law,

$$\oint_{\text{c.c.s.}} \mathbf{H} \cdot d\mathbf{l} = \int_S \mathbf{J} \cdot d\mathbf{S} + I$$

which relates the magnetic field intensity \mathbf{H} to the current, including in addition the displacement current density term $\frac{\partial \mathbf{D}}{\partial t}$, which was added

by Maxwell to make the law applicable to time-varying fields. The term \mathbf{J} represents the total current density. Equation (4) is the differential form of Faraday's law of electromagnetic induction. Equations (5), (6), and (7) are functional relationships, for the most part determined experimentally, by means of which the effects of different materials are accounted for. Mathematically, these equations are employed to reduce Eq. (3) and (4) to a pair of equations in only two unknowns. In free space, Eq. (5), (6) and (7) take their simplest form, respectively, $\mathbf{D} = \epsilon_0 \mathbf{E}$, $\mathbf{B} = \mu_0 \mathbf{H}$, $\mathbf{J} = 0$ (or $\mathbf{J} = \mathbf{J}_s$ a source current independent of \mathbf{E} and \mathbf{H}), where ϵ_0 and μ_0 are constants whose value depends on the system of units (in

the mks system $\epsilon_0 = 8.854 \times 10^{-12}$, $\mu_0 = 4\pi \times 10^{-7}$). Since matter itself is a relatively dilute collection of charged particles, it is always theoretically possible to define terms so that the theory is a description of the effects and interactions of charges in free space, with consequently no essential distinction between \mathbf{D} and \mathbf{E} or between, \mathbf{B} and \mathbf{H} , as indicated above. In practice however effects of materials are usually best handled in another way.^{6,7,9,10} Dielectric polarization effects are accounted for by making the \mathbf{D} vector include the electric dipole moment density \mathbf{P} , $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$, and then introducing a material constant, the permittivity ϵ , such that $\mathbf{D} = \epsilon \mathbf{E}$. The relative permittivity of a dielectric material is then equal to one plus the electric susceptibility. Magnetic polarization effects are handled similarly by defining the field vector \mathbf{B} so that it includes the magnetic dipole moment density \mathbf{M} , $\mathbf{B} = \mu_0 (\mathbf{H} + \mathbf{M})$. The material permeability is then introduced so that it depends upon the magnetic susceptibility analogously, and $\mathbf{B} = \mu \mathbf{H}$. Effects of conductors are represented by a material conductivity σ , such that $\mathbf{J}_c = \sigma \mathbf{E}$. With these simple forms for Eq. (5), (6) and (7), Eq. (3) and (4) take on the useful form

$$\nabla \times \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t} + \sigma \mathbf{E} + \mathbf{J}_s \quad (8)$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad (9)$$

provided μ and ϵ are constant in time. The term \mathbf{J}_s here includes currents arising from charges in free space plus any (source) currents which are independent of \mathbf{E} and \mathbf{H} . If there are no free charges in the region, \mathbf{J}_s includes only the source currents; these latter are known, so Eq. (8) and (9) may be solved for \mathbf{E} and \mathbf{H} . Since the equations are partial differential equations, boundary conditions over closed surfaces are required for unique solutions. Boundary conditions on the field quantities, which must hold at any boundary between two regions, may be derived from these equations. The conditions are: across a boundary (a) tangential \mathbf{E} must be continuous, (b) tangential \mathbf{H} must be continuous, (c) normal \mathbf{D} and normal \mathbf{B} must be continuous. Idealizations of material properties are sometimes helpful. For example, a perfect conductor has no non-static fields inside it, and at its surface, tangential \mathbf{E} and normal \mathbf{B} are zero, tangential \mathbf{H} is equal and perpendicular to any surface current density, and normal \mathbf{D} is equal to any surface charge density.

Two additional equations, especially useful in static problems, may be deduced from Eq. (2), (3) and (4):

$$\nabla \cdot \mathbf{D} = \rho \quad (10)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (11)$$

Solutions to the field equations are most readily obtained by imposing a restriction on the time dependence. If the fields are assumed to be independent of time (static), then Eq. (3) and (4)

or (8) and (9) decouple. One of the equations becomes $\nabla \times \mathbf{E} = 0$. This means that \mathbf{E} is irrotational and may be represented by a scalar potential function ϕ , $\mathbf{E} = -\nabla\phi$. Combining this with Eq. (10) gives the fundamental equation of electrostatics,

$$\nabla^2\phi = -\rho/\epsilon \quad (12)$$

Poisson's equation. This equation for the electrostatic potential is solved by the standard methods of partial differential equations. The boundary conditions on the potential may be found from the boundary conditions on the fields.^{9,10} In practice, it is frequently necessary to solve for the potential and electric field in a restricted region in which the charge density is zero, but the potential at the boundary is held at some particular value(s). The problem then is to solve Laplace's equation, $\nabla^2\phi = 0$, subject to the stated boundary conditions. The standard techniques for solving boundary value problems are employed. However, if the region of interest is partially open, known analytical techniques are sometimes inadequate to solve the problem. In two-dimensional problems of such a difficult type, the method of conformal transformations (conjugate functions) is often helpful.^{8,10}

The main applications of electrostatic theory are in (a) the theory of material properties, (b) the calculation of charged particle trajectories in electron guns, deflection systems, and accelerators (here in conjunction with magnetostatic theory), (c) the calculation of circuit component values, such as capacitance, and (d) the determination of voltage gradients in connection with voltage breakdown problems.

Magnetostatic theory is developed from Eq. (11) and (8). Since \mathbf{B} is divergenceless, it can be represented by the curl of a vector \mathbf{A} , which is known as the magnetic vector potential. Equation (8) can usually be written in terms of this potential as follows:

$$\nabla^2\mathbf{A} = -\mu\mathbf{J} \quad (13)$$

Taken one rectangular component at a time, this equation is of the same form as Poisson's equation [Eq. (12)] and may be solved in the same way. The boundary conditions on \mathbf{A} may be found from those on \mathbf{B} and \mathbf{H} . In regions with no current, Eq. (8) becomes $\nabla \times \mathbf{H} = 0$ so that \mathbf{H} may be represented by a scalar potential function $\mathbf{H} = -\nabla\phi_m$. In such regions then, in view of Eq. (11), the magnetic scalar potential, ϕ_m , must satisfy Laplace's equation

$$\nabla^2\phi_m = 0 \quad (14)$$

provided $\nabla\mu = 0$ in the region. The techniques and solutions of electrostatics are applicable to many magnetostatic problems. Unfortunately, however, in practice many of the systems designed to establish a given magnetic field incorporate ferromagnetic materials. For such materials, the magnetic susceptibility (and hence the permeability) is not independent of the field intensity and the field equations become nonlinear. Present mathematical techniques for

handling nonlinear problems are severely limited. Practical magnetostatic problems are, therefore, frequently solved by some approximation. One of the simplest and most useful approximations is a representation by a magnetic circuit.^{6,8,10} Series and parallel branches of the magnetic circuit may be recognized, and the techniques of linear and nonlinear circuit analysis can be applied to obtain a solution.

Magnetostatic theory is applicable to a myriad of magnetic devices, including deflection systems, motors, generators, relays, magnetic pickup devices, permanent magnets, memories, transducers and coils. To date, the need for particular solutions has frequently arisen before sound analytical methods have been available, so many devices are developed empirically.

Energy is required to establish electric and magnetic fields, and such energy is associated with the fields. The field energy in a given volume may be computed in most cases from a volume integral of one or both of the following energy density expressions $W_e = \frac{1}{2}\epsilon E^2$, $W_m = \frac{1}{2}\mu H^2$, respectively the electrostatic and magnetostatic values.

When the fields are time varying, Eq. (8) and (9) are coupled and must be solved simultaneously. Almost invariably, a potential function such as a vector potential or a Hertz potential is introduced.^{9,10} For example, Eq. (11) implies that \mathbf{B} may be replaced by a vector potential such that $\mathbf{B} = \nabla \times \mathbf{A}$. Equation (9) implies the following equation for \mathbf{E} ,

$$\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t} \quad (15)$$

so that \mathbf{H} and \mathbf{E} may be replaced in Eq. (8), and with the condition on \mathbf{A} , $\nabla \cdot \mathbf{A} = \mu\epsilon \partial\phi/\partial t$, the following equations may be obtained for \mathbf{A} and ϕ (σ assumed zero here)

$$\nabla^2\mathbf{A} = \mu\epsilon \frac{\partial^2\mathbf{A}}{\partial t^2} - \mu\mathbf{J} \quad (16)$$

$$\nabla^2\phi = \mu\epsilon \frac{\partial^2\phi}{\partial t^2} - \rho/\epsilon \quad (17)$$

That is, both the vector potential \mathbf{A} and the scalar potential ϕ satisfy a differential equation known as the inhomogeneous wave equation.

Because of their simplicity and practical importance, solutions for those sources and fields which simply oscillate at a single frequency have been studied extensively.^{2,9,10} In this case, the time is eliminated as an independent variable, as if by a transform operation. (In fact, transform methods are often the best means of obtaining transient field solutions.) In the equations, the time derivatives are replaced by frequency multipliers so that the resulting equations are functions of the space variables only. The vector potential may then be found by standard techniques of partial differential equations and boundary value problems. Having \mathbf{A} , the field quantity \mathbf{B} is found from $\mathbf{B} = \nabla \times \mathbf{A}$ and \mathbf{E} is found from

Eq. (8). In practice, a theorem which can be derived from the field equations, called the reciprocity theorem,^{2,3,10} is often helpful. The theorem relates the fields E_a and E_b produced respectively by a pair of current distributions J_a and J_b . The theorem is

$$\iiint E_a \cdot J_b dv = \iiint E_b \cdot J_a dv \quad (18)$$

For example, if J_b is selected to be a point current at point P , directed along x (represented mathematically by a Dirac delta function), then Equation (18), $E_{ax}(P) = \iiint E_b \cdot J_a dv$, gives a formula for the computation of the field due to J_a which is equivalent to a superposition integral involving a Greens function.

Perhaps the most fundamental problem of electromagnetic theory is the determination of the fields of a point charge, at rest, in oscillation, or in some general state of motion. For a point charge q , at rest in free space, the solution may be obtained by solving Eq. (12) in spherical coordinates. With the point charge at the origin, symmetry conditions may be employed to eliminate the angular variation, and the remaining differential equation in r can be solved subject

to Eq. (10) to give $\phi(r) = \frac{q}{4\pi\epsilon_0 r}$ for the potential associated with the point charge. A superposition integral

$$\phi = \iiint \frac{\rho dv}{4\pi\epsilon_0 r} \quad (19)$$

may then be employed to find the potentials associated with more complicated distributions. The field of an oscillating dipole, which is equivalent to a point alternating current, is also of great interest. This solution may be obtained from Eq. (16) (single frequency version). If the point current is directed along z , the z -component of the vector potential may be found by a procedure similar to that employed for a point charge. The final result is

$$A_{z0} = \frac{I\Delta z}{4\pi\mu_0 r} \cos \omega(t - \sqrt{\mu_0\epsilon_0}r) \quad (20)$$

where $I\Delta z$, the current moment, is equal to $q\Delta z$, the maximum dipole moment of the oscillating dipole. The factor $(t - \sqrt{\mu_0\epsilon_0}r)$ exhibits the time delay required for the effects of the oscillating charges to propagate to distant points. The electric and magnetic fields may be computed from Eq. (20) as indicated above. The magnetic field strength produced by an oscillating dipole (point current) is, for example, in the spherical coordinate system (r, θ, φ)

$$H_\varphi = \frac{I\Delta z}{4\pi} \sin \theta \left[\frac{\cos \omega(t - \sqrt{\mu_0\epsilon_0}r)}{r^2} - \frac{\omega \sqrt{\mu_0\epsilon_0}}{r} \sin \omega(t - \sqrt{\mu_0\epsilon_0}r) \right]$$

This form like Eq. (20), shows that the crests and valleys of the field oscillations are propagated in spherical waves at the speed of light $v = (\mu_0\epsilon_0)^{-1/2}$. The solution for a point current may be employed in an integral similar to Eq. (19) to find the vector potential of a more complicated distribution of current. Such solutions may also be employed to find the radiation patterns and input impedances of antennas.

The potentials and fields produced by a charge moving in an arbitrary way may also be obtained.^{3,9} The results may be found in Stratton, 9, pp. 475-476.

In regions free of source currents and charges, the fields and potential satisfy the homogeneous wave equation [for example Eq. (16) with $J = 0$]. Then one of the simpler solutions which can be obtained is that of the plane electromagnetic wave. With appropriate orientation of the rectangular coordinate system, the solutions show that plane waves may progress along z , with components as follows:

$$E_r = E_0 \cos \omega(t - \sqrt{\mu_0\epsilon_0}z)$$

$$H_\varphi = E_0 \sqrt{\frac{\epsilon_0}{\mu_0}} \cos \omega(t - \sqrt{\mu_0\epsilon_0}z)$$

where E_0 is an arbitrary constant amplitude. Note that E , H and the direction of propagation are all perpendicular to one another. The Poynting vector, $S = E \times H$, points in the direction of propagation. Moreover, the power carried through a closed surface by an electromagnetic field may be computed from a surface integral of the Poynting vector.

With single frequency fields in source free regions, both H and E can be represented by vector potentials,^{2,9,10} $H_1 = \nabla \times A_1$, $E_2 = \nabla \times A_2$, and moreover the coordinate systems may be oriented so that A_1 and A_2 each have a single component¹⁰. In cylindrical systems, this single component is commonly along z . H_1 is then transverse to z (TM) and the set of fields, E_1 , H_1 , derivable from A_1 , are called TM fields. E_2 is likewise transverse to z and the set of fields, E_2 , H_2 , derivable from A_2 , are called TE fields. This procedure is particularly helpful in problems involving transmission lines and waveguides and is developed in detail in Weeks¹⁰, ch. 4-6.

Some of the most interesting and fundamental problems of electromagnetic theory are concerned with the scattering and diffraction of electromagnetic waves.^{2,3,8-10} For example, exact solutions are available for the scattering by cylinders and spheres, as well as an infinitely long slit. Approximate solutions are available for many other shapes. The methods are those outlined above, supplemented by generalizations of the principles of Huygens and Babinet.

Another topic of wide interest is the nature of fields in ionized gases or plasmas. The applications range from ionospheric propagation to microwave devices to nuclear apparatus to magneto-hydrodynamics to satellite reentry problems.

The simplest theory for these effects is developed from Eq. (3) and (4) (single frequency version) by separating the ion current term $J_e - \rho v$ from J , and employing Newton's law to eliminate v in favor of E , H and whatever mechanical constraints are applicable^{4,9,10} (see PLASMAS).

Effects peculiar to charges moving with very high velocities have not been included in this discussion (see RELATIVITY THEORY). Quantum effects are also discussed elsewhere (see QUANTUM ELECTRODYNAMICS and QUANTUM THEORY).

W. L. WEEKS

References

1. Hayt, W. H., "Engineering Electromagnetics," New York, McGraw-Hill Book Co., 1958.
2. Harrington, R. F., "Time-Harmonic Electromagnetic Fields," New York, McGraw-Hill Book Co., 1961.
3. Jackson, J. D., "Classical Electrodynamics," New York, John Wiley & Sons, 1962.
4. Javid, M., and Brown, M., "Field Analysis and Electromagnetics," New York, McGraw-Hill Book Co., 1963.
5. Panofsky, W., and Phillips, M., "Classical Electricity and Magnetism," Reading, Mass., Addison-Wesley, 1955.
6. Peck, E. R., "Electricity and Magnetism," New York, McGraw-Hill Book Co., 1953.
7. Plonsey, R., and Collin, R., "Principles and Applications of Electromagnetic Fields," New York, McGraw-Hill Book Co., 1962.
8. Smythe, W. R., "Static and Dynamic Electricity," Second edition, New York, McGraw-Hill Book Company, 1950.
9. Stratton, J. A., "Electromagnetic Theory," New York, McGraw-Hill Book Co., 1941.
10. Weeks, W. L., "Electromagnetic Theory for Engineering Applications," New York, John Wiley & Sons, 1964.

Cross-references: ELECTRICITY, QUANTUM ELECTRODYNAMICS, PLASMAS, POTENTIAL, QUANTUM THEORY, RELATIVITY, STATIC ELECTRICITY, WAVEGUIDES.

ELECTRON

The electron is the smallest known electrically charged particle. Its existence and characteristics were inferred from many experiments clustered in and around the last decade of the nineteenth century. In the 1830's, Faraday had tentatively suggested that his experiments in ELECTRO-CHEMISTRY could be interpreted in terms of a small unit of charge attached to ions. This notion of individual "atoms of charge" was somewhat eclipsed, however, by the enormous success of Maxwell's theory of electromagnetism, which was generally interpreted, by 1880, as favoring a view that electrical phenomena were due to continuous charge distributions and motions. G. Johnstone Stoney, in 1874, and Helmholtz, in 1881, had suggested again an atomic interpretation of electricity, but it was not until the brilliant

experiments of Perrin, J. J. Thomson, Zeeman, and others in the 1890's that the concept of the electron received firm experimental foundation. Later experiments and theory (Millikan, Bohr, etc.) established the constancy of the electronic charge and interwove the concept of an electron of definite charge and mass into the basic structure of the atom.

The Cathode Ray Controversy. After the discovery of the cathode ray in high-vacuum discharge tubes by Plücker in 1858, there developed, with the experiments of Goldstein, Crookes, Hertz, Lenard, and Schuster, a controversy over the nature of the rays. A predominately German school held that the rays were a peculiar form of electromagnetic rays. The British physicists thought they were negatively charged particles. The controversy provides a classic "case history" of the typical scientific controversy in which two quite different models both explain most, but not all, of the observable facts. The proponents of each model designed ingenious experiments and in some cases were so trapped in their preconceptions that they badly misinterpreted their observations. The Germans were especially impressed by the fact that the rays could go through thin foils—something no known particles could do. The British were firm in pointing out that the rays could be deflected by magnetic fields—something not possible with electromagnetic waves. Hertz, in what he thought was a crucial experiment, was unable to detect deflection of the rays by electric fields, but this very phenomenon was demonstrated by J. J. Thomson and made the basis for his conclusive experiments that the rays had velocities less than that of light. Thomson showed, further, that if one assumed that the rays were composed of particles, then the particles had the same ratio of charge to mass regardless of the cathode material or the nature of the residual gas. Perrin's classic experiment, meanwhile, proved that the rays did indeed convey negative charge. In the decade between 1896 and 1906, Thomson and others showed that negatively charged particles from sources other than cathode rays had the same ratio of charge to mass: the negative particles emitted by hot filaments in the Edison effect, the beta rays emitted by some radioactive materials, and the negative particles emitted in the photoelectric effect that had so ironically been discovered by Heinrich Hertz in his great experiment which demonstrated the electromagnetic rays predicted by Maxwell's equations.

Thomson's Determination of e/m . In 1897 Thomson devised an apparatus in which he could deflect a beam of cathode rays with a magnetic field of induction B and also with an electric field of strength E . If the fields are perpendicular to each other, and to the original path of the beam, and if they occupy the same region, then (with proper polarities and magnitudes of fields) the electric force on the beam can equal the magnetic force, so that the beam hits the same point on a fluorescent screen as when no fields are applied. If e is the charge of a

given particle, m its mass, and v its velocity, $v = E/B$. Thus, velocities of typical cathode ray beams could be measured. If the magnetic field is used alone, and the curvature R of the beam is measured, then one can equate centripetal and magnetic field forces $mv^2/R = Bev$, and then deduce $e/m = v/BR$. With v known from the previous experiment, e/m can be calculated. Thomson's early values were not very precise, but later experiments of a similar type gave values close to 1.76×10^{11} coulombs/kg. More recent experiments, drawing on measurements of many kinds, give $e/m = (1.75890 \pm 0.00002) \times 10^{11}$ coulombs/kg.

The Zeeman Effect. In 1896 Zeeman discovered the broadening of spectral lines when a light source was in a strong magnetic field. Experimental refinements by Zeeman and others, and theoretical work by Lorentz and Zeeman, permitted the interpretation of this effect as due to the influence of the magnetic field on oscillating or orbiting negatively charged particles within the light-emitting or absorbing atoms. From the spectroscopic data, the ratio of charge to mass of these hypothetical particles could be shown to be equal to that of cathode rays. The Zeeman effect thus provided the first experimental evidence that the negative particles emitted by atoms when heated (Edison effect) or subject to high fields and/or ionic bombardment (cathode rays) or bombarded by short-wavelength light (photoelectric effect) were, indeed, actual constituents of the atoms and were probably responsible for the emission and absorption of light.

The Charge on the Electron. In the decade following 1897, many different methods were evolved for determination of ionic charges. Some methods depended upon measuring the total charge of a number of ions used as nuclei for cloud droplet formation. Other methods were more indirect—experiments, for example, which, combined with the kinetic theory of gases, could give crude values for avogadro's number, N (see MOLE CONCEPT). By dividing the Faraday constant (the charge carried in electrolysis by ions formed from one gram-atom of a univalent element) by N , one could determine the average charge per ion. Similarly, the constants in Planck's theory of blackbody radiation, when evaluated experimentally, could provide a numerical value for N , as could certain experiments in radioactivity. All such methods gave values of N of the order of 6×10^{23} , and hence 1.6×10^{-19} coulomb for the ionic charge. None of these methods measured individual charges; strictly speaking, the value for the ionic charge could be thought of only as an average value.

Millikan's experiments with single oil drops, beginning in 1906, provided a method for measuring extremely small charges with precision. He was able to show that the charge on his drops was *always* ne , with $e = 1.60 \times 10^{-19}$ coulomb (modern value) and n a positive or negative integer.

He observed the motions of very small charged oil drops in uniform vertical electric fields. The

drops were so small that they moved with constant velocity (except for Brownian fluctuations) for a given force. The force in each case was due to gravity acting on the mass of the drop and to the electric field (if any) acting on the charge, q , on the drop. The charge on a given drop could be changed by shining x-rays upon it. Using Stokes' Law, in a form modified to correct for the fact that the drops were *not* large in comparison to the inhomogeneities of the surrounding air, and the velocity of a drop in free (gravitational) fall, Millikan could infer the diameter and mass of a given drop, and then calculate its charge. The charge q always equaled ne . (See reference 1 or 2 for experimental details.) A few other physicists, in similar experiments, thought they had detected electric charges smaller than Millikan's e , but their experimental techniques were probably faulty.

Millikan's experiment did not prove, of course, that the charge on the cathode ray, beta ray, photoelectric, or Zeeman particle was e . But if we call all such particles electrons, and assume that they have $e/m = 1.76 \times 10^{11}$ coulombs/kg, and $e = 1.60 \times 10^{-19}$ coulomb (and hence $m = 9.1 \times 10^{-31}$ kg), we find that they fit very well into Bohr's theory of the hydrogen atom and successive, more comprehensive atomic theories, into Richardson's equations for thermionic emission, into Fermi's theory of beta decay, and so on. In other words, a whole web of modern theory and experiment defines the electron. (The best current value of $e = (1.60206 \pm 0.00003) \times 10^{-19}$ coulomb.)

The Wave Nature of the Electron. In 1924, L. DeBroglie suggested that the behavior of electrons within atoms could be better understood if it were assumed that the motion of an electron depends upon some sort of accompanying wave, the length of which would be h/p (h = Planck's constant and p the momentum of the electron). This suggestion led to the development of QUANTUM MECHANICS by Schrödinger, Heisenberg, and others. The concept of electron waves provided an explanation for experiments on reflection of electron beams by metallic crystals, carried out from 1921 onward by Davisson and others, and provided an impetus for the experiments of G. P. Thomson on the diffraction of electron beams by thin films (see ELECTRON DIFFRACTION).

Other Characteristics of Electrons. In applying quantum mechanics to certain problems in atomic spectroscopy, in 1925 and 1926, Pauli, and Goudsmit and Uhlenbeck found that electrons must possess angular momentum of amount $\pm \frac{1}{2}(h/2\pi)$. Dirac's work on a generalized quantum theory of the electron showed that it possessed a related magnetic dipole moment of magnitude $eh/4\pi mc$ (see ELECTRON SPIN). The ratio of the dipole moment to the angular momentum (e/mc) is larger than can be accounted for in classical terms with any homogeneous wholly negative model. The concept of electronic dipole magnetic moment is essential not only in spectroscopy but in theories of ferromagnetism (see MAGNETISM).

One may speak of the "classical radius of the electron," $a = e^2/mc^2$, derived by setting the self-energy of the coulomb field of a charge e contained at a radius a equal to the relativistic rest energy, mc^2 of the electron. This $a = 2.82 \times 10^{-13}$ cm, comfortably smaller than any atom, but larger than the usual estimates of sizes of protons and neutrons.

Positive Electrons. Dirac's paper in 1928 could be interpreted as predicting the existence of electrons that are positive. But until such particles were found experimentally by C. D. Anderson in 1932 in cloud chamber pictures of cosmic ray particle tracks, most physicists preferred other interpretations of Dirac's paper. Positive electrons, or **POSITRONS** are now known (1) to occur as decay products from certain radioactive isotopes, (2) to be produced (paired with a negative electron) in certain interactions of high-energy gamma rays with intense electric fields near nuclei, and (3) to be the product of certain decays of certain mesons. In principle, positrons could form anti-atoms with nuclei made from anti-protons and anti-neutrons, but in practice almost all positrons produced in the observable universe quickly meet their end by annihilating themselves together with some hapless negative electron. The end product of a positron-electron annihilation is a pair of gamma rays.

Applications of Electrons. Aside from their inherent usefulness in physical theories of magnetic, electrical, optical, and mechanical properties of matter, electrons either in beams or in conductors can be made to do all sorts of useful things. Cathode ray oscilloscopes, electron microscopes, image converters, certain memory devices for computers, television picture tubes, and most "radio tubes" depend upon beams of electrons controlled by electric or magnetic fields (see **ELECTRON OPTICS**). In ordinary metallic conductors, electricity is carried primarily by electrons. The behavior of electrons in **SEMI-CONDUCTORS** and in **superconductors** (see **SUPER-CONDUCTIVITY**) has in recent years been the basis both of intense theoretical interest and of interesting and useful devices.

DAVID L. ANDERSON

References

1. Millikan, R. A., "The Electron," edited with an introduction by J. W. M. DuMond, Phoenix Science Series, PSS523, University of Chicago Press, 1963.
2. Anderson, D. L., "The Discovery of the Electron," Princeton, N.J., D. Van Nostrand Co., 1964.
3. Shankland, R. S., "Atomic and Nuclear Physics," Second edition, New York, The Macmillan Company, 1960.
4. Condon, E. U., and Odishaw, H., "Handbook of Physics," pp. 7-169, New York, McGraw-Hill Book Co., Inc., 1958.

Cross-references: ELECTROCHEMISTRY, ELECTRON DIFFRACTION, ELECTRON OPTICS, ELECTRON SPIN, MAGNETISM, MOLE CONCEPT, PHOTOELECTRICITY, POSITRON, QUANTUM MECHANICS, SUPERCONDUCTIVITY, ZEEMAN AND STARK EFFECTS.

ELECTRON DIFFRACTION

The discovery of electron diffraction independently by C. J. Davisson and L. H. Germer (1927) and G. P. Thomson (1927) verified L. de Broglie's earlier hypothesis (1924) that matter exhibits both corpuscular and wavelike characteristics. This hypothesis served as a stimulus for the formal development of quantum mechanics by E. Schrödinger, M. Born, W. Heisenberg, and others. Following this momentous discovery, which eventually resulted in the award of a Nobel Prize to Davisson and Thomson, electron diffraction was immediately utilized as a tool for the study of the structure of matter.

Electron, x-ray and neutron diffraction are all used for structure studies. Electron diffraction is used particularly for those structural studies that involve small numbers of atoms. This is due to the strong interaction of electrons with matter. Thus the principal area of application of electron diffraction is for the study of thin films, surfaces, gases and small samples.

The different energy ranges that were used in the Davisson-Germer and Thomson experiments provide a natural division for a description of the types of equipment, areas of application, and analytical techniques that have evolved since 1927. These experiments were performed with electrons having energies in the vicinity of 150 eV and 15 keV, respectively. De Broglie's relationship $\lambda = h/mv$, where h is Planck's constant and λ is the wavelength associated with a mass m traveling with a group velocity v , reduces to $\lambda \approx \sqrt{150/V}$ for electrons in the nonrelativistic limit, where V is the accelerating voltage and λ is expressed in Angstroms. The two experiments thus used electrons having a wavelength of approximately 1 and 0.1 Å. The longer wave-length is comparable to the spacing between atoms in crystals.

While 50-keV electrons, which are used in most commercial electron diffraction instruments, penetrate to a depth of about 10^3 Å into a crystal, 150-eV electrons penetrate only about 10 Å. Since the higher-energy electrons are capable of passing through the several layers of adsorbed foreign material that normally are present on the surface of a crystal, surface cleanliness and therefore the vacuum requirements for 50-keV electron diffraction are not as stringent as those for low-energy electron diffraction. This factor, in addition to the relative ease in focusing intense high-voltage beams, and individual interests, resulted in the wide application of high-energy electron diffraction for structure studies. A typical instrument of this type operates at about 50 keV, has provisions for producing and focusing the electrons, contains specimen manipulators, photographic means to record the diffraction patterns and is contained in a chamber capable of being evacuated to 10^{-5} torr. Diffraction patterns are obtained either by transmission of the beam through very thin films or by working at grazing incidence and reflection. The grazing incidence technique is potentially capable of resolving structures having a monolayer thickness.

One of the most common uses of 50-keV electron diffraction by transmission or reflection is for the study of films on amorphous, polycrystalline and single crystal substrates. This includes films that have been formed by the oxidation or corrosion of a surface, as well as those formed by the deposition of material on a substrate. In many instances these films consist of crystallites having an orientation that is related to the structure and orientation of the substrate material.

In the past few years, the transmission ELECTRON MICROSCOPE has become widely used for the study of atomic arrangements at lattice imperfections, such as dislocations and stacking faults. Image contrast is obtained by local differences in the intensities of diffracted beams. In addition, many electron microscopes are constructed in such a way that it is possible to obtain the diffraction pattern associated with the material in the area being studied.

Electron diffraction in this energy range is especially useful for the determination of the atomic arrangements, bond distances, bond angles, and mean square atomic vibrational amplitudes in gaseous molecules.

Low-energy electron diffraction was used by only a few groups until approximately 1960. Improved diffraction equipment, which enabled the direct display of the diffraction pattern on a fluorescent screen by accelerating the diffracted electrons after they had passed through grids, and the commercial availability of ultrahigh (10^{-10} torr) vacuum equipment, resulted in a resurgence of interest in this field. The structure of clean surfaces, the arrangements of foreign atoms on these surfaces at a monolayer or less coverage, and many aspects of the initial stages in the oriented overgrowth of thin films have been studied with electrons having energies in the range of 2 to 10^3 eV. It has been revealed that the atomic arrangement at the clean surfaces of semiconductors such as germanium and silicon is quite unlike that found in the bulk of these materials. At low coverages, foreign atoms are normally adsorbed in structures that have a symmetry and dimensions that are simply related to the orientation of the substrate plane. A multitude of such structures has been found on semiconductors and metals. Their atomic array is dependent on many parameters, such as the amount of adsorbed material, the temperature, orientation and cleanliness of the substrate.

Fundamental properties of solids, such as characteristic energy losses, atomic mean square vibrational amplitudes and electron range, have also been investigated by both high- and low-energy electron diffraction.

ALFRED U. MAC RAE

References

- Pinsker, Z. G., "Electron Diffraction," London, Butterworth's Scientific Publications, 1953.
Heidenreich, R. D., "Fundamentals of Transmission Electron Microscopy," New York, John Wiley & Sons, 1964.

Cross-references: DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS, ELECTRON, ELECTRON MICROSCOPE, ELECTRON OPTICS.

ELECTRON MICROSCOPE

The electron microscope is a device that forms magnified images by means of electrons. The electrons are usually accelerated to between 50 and 100 kV. The microscope magnifies in two or three stages by means of electromagnetic or electrostatic lenses.

In 1878 Ernst Abbé proved that the resolution of the optical microscope is limited by the wavelength of light. No matter how perfect and free of aberrations the optical system is, the image of a geometrical point is not a point but a disc, the "Airy disc." Regardless of any further magnification, two separate points cannot be resolved as separate unless their centers are the distance d apart, whereby d is the diameter of the Airy disc referred to the object plane:

$$d = \frac{0.5 \lambda}{n \sin \alpha} \quad (1)$$

where

- λ = wavelength of the illuminating light
- n = index of refraction of medium between object and lens (in air, $n = 1$)
- α = aperture of lens, i.e., half-angle of collected light beam
- $n \sin \alpha$ = NA = numerical aperture of lens (in air, $NA_{\max} = 0.95$).

Even if we go to the extreme of using the ultraviolet line of mercury ($\lambda = 253.7$ nm*), oil immersion optics ($NA = 1.4$), quartz lenses and microphotography, the best resolution obtainable is still:

$$d = 110 \text{ nm}$$

The electron microscope makes use of the wave properties of the moving electron. Its "de Broglie" wavelength is:

$$\lambda = \frac{h}{mv} = \frac{1.23}{V^{1/2}} \text{ [nm] for nonrelativistic electrons (below 20 kV)} \quad (2a)$$

and

$$\lambda = \frac{1.23}{(V + 10^{-6} V^2)^{1/2}} \text{ [nm] for relativistic electrons} \quad (2b)$$

where

- h = Planck's constant
- m = mass of electron
- v = velocity of electron
- V = accelerating voltage in volts.

For electrons of $V = 50$ kV the wavelength is therefore:

$$\lambda_{50 \text{ kV}} = 0.00535 \text{ nm}$$

* 1 nanometer (nm) = 1 millimicron (m μ) = 10 angstrom units (Å).

Objective lenses for electrons, unlike those for light optics cannot be made free of spherical aberrations. They have to operate with numerical apertures that are 500 to 1000 times lower. The best theoretical resolution for magnetic lenses working with electrons of 50 kV energy is:¹

$$\delta = 0.21 \text{ nm}$$

The best practical resolution is between 0.5 and 1 nm.

The first electron microscope was built by M. Knoll and E. Ruska at the Technical University of Berlin early in 1931. It had two electromagnetic lenses in series and achieved a modest magnification of 17. Knoll and Ruska improved the electron microscope step by step. They added a condenser lens and built an iron shield with a narrow center gap around their magnetic lens.

Ruska, working from 1932 by himself, equipped the magnetic lenses with narrow pole pieces and was able to demonstrate in 1933 a resolution of 50 nm (magnification of 12,000), better than the best obtainable optical resolution. Figure 1 shows a functional diagram of his supermicroscope. Figure 2 shows the details of one of his lenses.

Parallel to the development of the magnetic electron microscope went the development of an electrostatic one. In 1931, Brüche and Johansson of the Research Institute of A.E.G. of Berlin imaged the emitting surface of the cathode with an electrostatic immersion objective. In 1932, they employed unipotential or Finzel-lenses used now, almost exclusively, in general-purpose electrostatic microscopes.

Magnetic Lenses. A charged particle entering a uniform magnetic field perpendicular to the

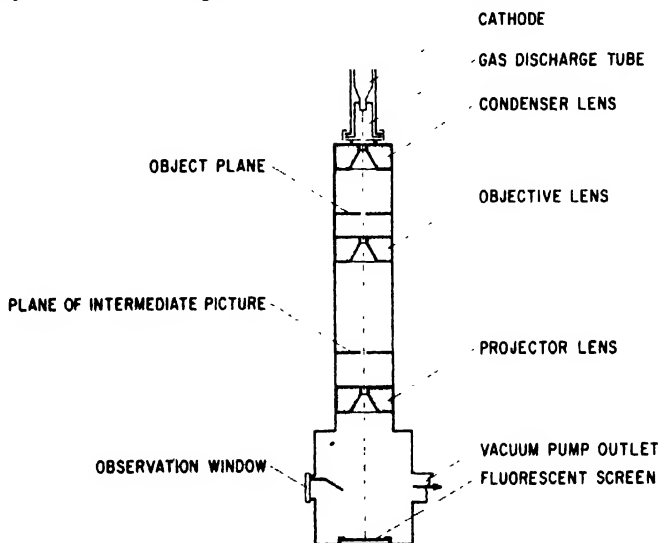


FIG. 1. First Supermicroscope (Ruska 1933).

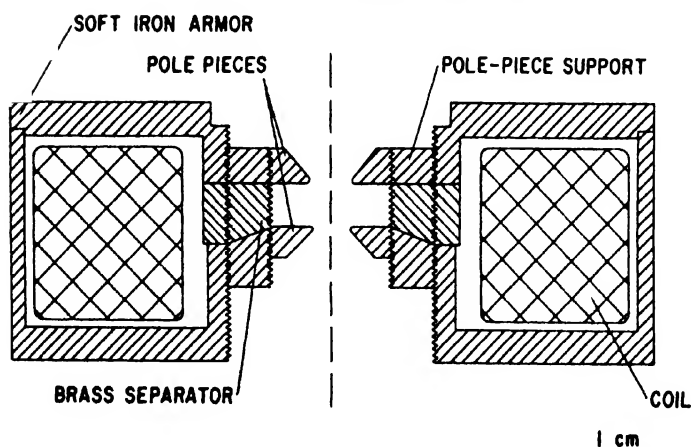


FIG. 2. Magnetic Lens (Ruska 1933).

lines of force will describe a circle; moving parallel to them, it will not be deflected. The radius of the circle that the particle describes, the "cyclotron radius", is:

$$\rho = \frac{m v}{e B} \quad (3)$$

where

m = mass of particle
 e = charge of particle
 v = velocity
 B = magnetic field intensity (gauss).

The circle described by an electron is

$$\rho_e = 3.372 \frac{\sqrt{V}}{B} \text{ [cm] for nonrelativistic electrons} \quad (4a)$$

$$\rho_e = \frac{1}{B} \sqrt{11.3V + 1.11 \times 10^{-5} V^2} \text{ [cm] for relativistic electrons} \quad (4b)$$

where V is in volts and B is in gauss.

The time it takes a particle to describe a cyclotron circle is:

$$\tau = \frac{2\pi\rho}{v} = 2\pi \frac{m}{e B} \text{ [sec]} \quad (5)$$

A charged particle entering a uniform magnetic field at an angle will describe a cycloid. While its velocity component normal to the field lines causes it to describe a circle, the velocity component parallel to the field lines remains unchanged. Since the time required to describe a circle is independent of the normal velocity, all circles, large or small, are traversed in the same time interval. In an electron beam with low divergence where, therefore, the velocities parallel to the magnetic field (v_z) are identical, all electrons leaving one point P on the axis will meet downstream at another point P' at a distance

$$d = \frac{v_z}{\tau} \quad (6)$$

with angles to the axis identical to those they had at point P (see Fig. 3).

A "long" magnetic lens, having a uniform magnetic field extending from the object to the image point, will form an upright picture of the object with an image to object ratio of 1 to 1.

A "thin" magnetic lens is a lens with a magnetic field short compared to the object-to-image distance. A "weak" thin lens is a lens with the focal length large compared to the axial length of the magnetic field. The refractive power, i.e., the reciprocal of the focal length, of such a lens is determined by:

$$\frac{1}{f} = \frac{0.022}{V} \int_{-\infty}^{+\infty} B_z^2 dz \text{ [cm}^{-1}] \quad (7)$$

It images according to the general optical equation:

$$\frac{1}{f} = \frac{1}{a} + \frac{1}{b} \quad (8)$$

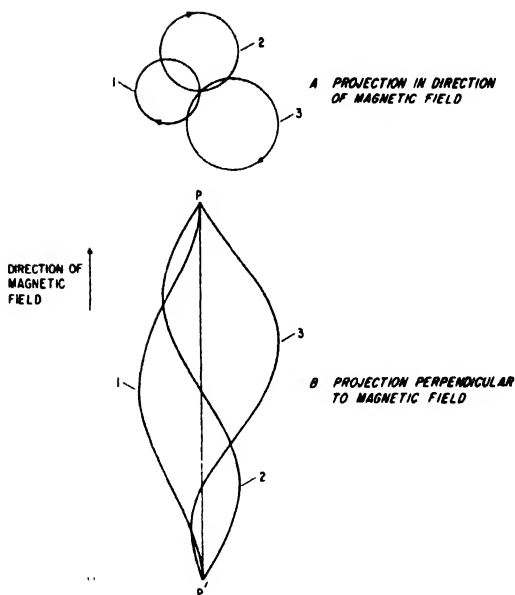


FIG. 3. Cycloids, described by electrons in uniform magnetic field.

where a is the object distance and b is the image distance. The picture is turned around from the position of the object. The angle it is turned is determined by:

$$\theta_i = \frac{0.149}{\sqrt{V}} \int_{-\infty}^{+\infty} B_z dz \text{ [radians]} \quad (9)$$

The objective and projector lenses of the electron microscope require extremely short focal lengths in order to obtain high magnifications without going to extremely long microscopes. These lenses have, therefore, pole pieces which limit the extent of the magnetic field, both in the axial and radial dimensions. The treatment of strong, thin magnetic lenses can be found in references 2 and 3.

Electrostatic Lenses. *Immersion lenses* consist of two apertures or two coaxial cylinders at different potentials. They are important as lenses for television or oscillograph cathode-ray tubes. The electrostatic lenses used exclusively as objective or as projector lenses in electron microscopes are *unipotential* or *Einzel-lenses*. They have the same potential on either side of the lens. They consist of three apertures: the two outer apertures are at ground or anode potential, and the center electrode can have either a positive or negative potential. Regardless of whether a positive or a negative potential is applied, the lens will always be convergent. Figure 4(a) shows the equipotential lines of such a lens; Fig. 4(b) shows the focal length vs V_L/V_0 for this lens, where V_L is the voltage applied to center electrode and V_0 the cathode potential. Figure 5 shows the design of a typical electrostatic lens.

Main Features of Transmission Microscope. The normal electron microscope is a transmission

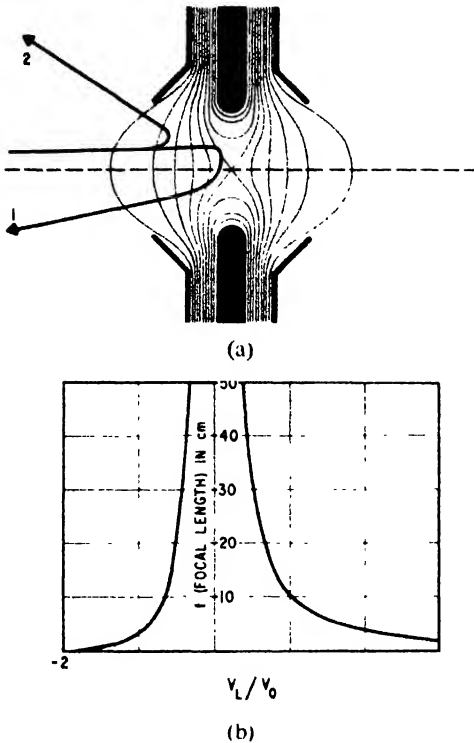


FIG. 4. Unipotential or Einzel-lens:
(a) Equipotential Lines (From H. Mahl and A. Pendzich, *Zeitscher. Techn. Phys.*, **24**, 38-42 (1943))
(b) Focal Length vs V_L/V_0 (From H. Johannson and O. Scherzer, *Zeitscher Physik*, **80**, 183-202 (1933))

microscope where the image is formed by electrons which have passed through the specimen. It is composed of the following major sections: electron gun, condenser lens, specimen chamber, objective lens, projector lens, viewing and photographing chamber. It has the following ancillary equipment: (1) power supplies for high voltage, heater voltage, and focusing currents; (2) vacuum systems.

The *electron gun* generates the electron beam which illuminates the object. It has to provide the required electron density within a certain limited divergence. It consists of a hairpin tungsten filament enclosed in a cup-shaped electrode at cathode or a more negative potential. While the anode facing the cathode is at ground potential, the cathode is maintained at a high negative potential, usually between 50 and 100 kV.

The *condenser lens* or lenses increase the electron density reaching the specimen by concentrating the beam. An aperture in the condenser lens of 0.25 to 0.5 mm diameter reduces the amount of stray electrons reaching the specimen.

The *objective lens* provides the first magnification. It is a strong, thin lens with a high refractive power. It has, therefore, narrow precision pole pieces. The object is brought very close to the magnetic gap.

The different shades of density of the object are reproduced by the scattering of electrons. The electrons removed from the image cone of each point will reach the image plane at random. They will, therefore, produce fogging of the image. In order to remove as many stray electrons as possible, an aperture of 25 to 100 μ is inserted in the objective gap.

The *projector* or *image lens* selects a small portion of the intermediate image produced by the objective lens and magnifies it again. A third electron-optical magnification is in some cases produced by a second projector lens.

The *specimen chamber* is located above the objective lens. It can be opened for inserting the specimen, usually without disturbing the vacuum in the main column. It has adjustments that permit shifting the specimen in the object plane in order to locate the area of interest. Some models have special facilities to keep the specimen at certain high or low temperatures.

The *viewing chamber* at the bottom of the microscope column contains the fine-grain fluorescent screen that can be observed through glass windows either directly or through a monocular or binocular telescope. If a permanent record is desired, the fluorescent screen is moved aside and a photographic plate inside a plate holder is exposed. The electrons produce a latent image

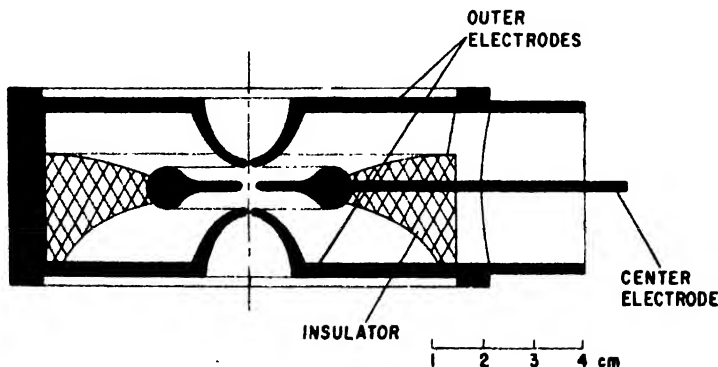


FIG. 5. Typical Electrostatic Lens. Based on article by H. Mahl, *Jahrb. AEG Forschung.*, **7**, 43-56 (1940).

directly. Each plate can be removed from the vacuum separately.

In some cases when an extremely high magnification, for instance 100,000 times, is desired, it may pay to magnify electron-optically to a somewhat lower magnification and to add a final photographic enlargement later; the adjustment of the electron microscope is then much easier, since the field of vision is so much larger.

A photographic film is used sometimes, instead of plates, if a series of pictures is to be taken and if extremely high resolution is not required.

The *vacuum system*, generally maintaining a vacuum of 10^{-4} to 10^{-5} torr (1 torr = 1 mm Hg), consists usually of an oil diffusion pump backed by a mechanical forepump. A second mechanical pump is sometimes used to purge the specimen chamber and the photographic plate lock before they are opened to the main vacuum column. A water or liquid-nitrogen-cooled baffle is used to prevent backstreaming of oil vapor into the chamber.

Power Supplies. The voltage regulation required for the power supplies of the electrostatic microscope is not very critical. As long as the lens voltage and cathode potential maintain the same linear ratio—very often they are identical—a good image is obtained.

The refractive power of the magnetic lens depends directly on the square of the magnetic flux, which is proportional to the lens current if the magnetic circuit is not saturated, and varies inversely with the cathode potential [see Eq. (7)].

It is, therefore, paramount that all power supplies be extremely well-regulated. Assuming a maximum permissible unsharpness of 1 nm, Zworykin *et al.*¹ give the following values for maximum permissible instabilities:

$$\text{High voltage power supply: } \frac{\Delta V}{V} = 1.1 \times 10^{-4}$$

$$\text{Objective lens: } \frac{\Delta I_1}{I_1} = 0.55 \times 10^{-4}$$

$$\text{Projector lens: } \frac{\Delta I_2}{I_2} = 1.3 \times 10^{-4}$$

$$\text{Condenser lens: } \frac{\Delta I_3}{I_3} = 1 \times 10^{-4}$$

The values guaranteed by various manufacturers are often considerably lower, especially during the short exposure time.

Other Types of Electron Microscopes. Several types of electron microscopes that are different from the transmission microscope have been described, but none of them has attained its popularity.

The *emission electron microscope* has been used to study various surface phenomena. The *thermionic emission microscope* is almost as old as the transmission microscope. The first electrostatic microscope and one of the first electromagnetic microscopes were used to study electron

emission from hot cathodes. Secondary emission and photoemission have also been studied. For the *field emission microscope*, see FIELD EMISSION.

The *shadow microscope* of Boersch^{5,6} is an inverted microscope. Two electrostatic lenses with large object and small image distances reduce the size of the crossover point (the smallest beam diameter located between grid and anode) by several orders of magnitude. This fine spot is the illuminating source for the object, which is brought close to the image plane of the second lens. The magnification depends on the ratio of the spot to object and the spot to screen distances. No lenses are required between the object and the screen. The ultimate resolution depends only on the size of the spot. The resolution could be of the same order as that of the transmission microscope. The spot-to-object distance is very critical, but it is difficult to reproduce. It is, therefore, difficult to measure the magnification.

The *scanning microscope*, first proposed by M. von Ardenne in 1938, also uses an inverted microscope. In this case, the fine spot is deflected in a raster, similar to the television scan, across the object. The deflection of the beam takes place between the first and the second reducing lens. In this manner, it is not necessary to introduce a deflection system between the last lens and the object which would increase the image distance considerably. A recording drum is mounted below the specimen, which makes one revolution per scanning line. The paper speed is 2500 times the speed of the spot across the specimen. This ratio is, therefore, the image magnification. To increase the signal-to-noise ratio, a slow scan is used. It takes a period of 10 minutes or more to record one picture.

Another scanning microscope was built by Zworykin and co-workers.⁷ It uses a collector above the specimen to collect secondary electrons emitted from the surface of an opaque object. The collected electron current is amplified and modulates the electron beam current of a display tube whose x and y deflection are synchronized with the microscope beam scan.

Mirror Microscope.⁸ If we were to apply to the unipotential lens in Fig. 4(a) a potential much more negative than that of the cathode, a zero potential line would cross the center of the lens, which the electrons cannot penetrate. They would have to turn back. The unipotential lens is then turned into a mirror. If the electrons are able to penetrate into the convergent center section of the lens, the mirror will act as a convergent mirror [electron path 1 in Fig. 4(a)]. If the potential of the center electrode is so highly negative that the electrons can penetrate only into the divergent outer section, the mirror is divergent (electron path 2).

If the center electrode of the unipotential lens is closed and is kept at cathode or a somewhat more negative than cathode potential, the unipotential lens would be transformed into a divergent immersion mirror. By keeping the potential of this electrode very close to cathode potential, the electrons will penetrate close to

its surface before they turn around. At the lowest point, they will have only a very small tangential energy. They are, therefore, easily influenced by various parameters of the metal surface. e.g., by weak magnetic fields, by electrostatic charges or by distortions in the electrostatic field close to the surface due to surface roughness. They form, therefore, an enlarged image of certain surface features if a fluorescent screen is mounted some distance away from the mirror, e.g., on the underside of the preceding lens. This screen will always have a small aperture in order to admit the electron beam to the mirror.

MARTIN M. FREUNDLICH

References

1. Ruska, E., "Fifth International Congress for Electron Microscopy," New York, Academic Press, 1962.
2. Ruska, E., *Arch. Elektrotechn.*, **38**, 102-130 (1944).
3. Hall, C. E., "Introduction to Electron Microscopy," Chapter 5, New York, McGraw-Hill Book Co., 1953.
4. Zworykin, V. K., Morton, G. A., Ramberg, E. G., Hillier, J., and Vance, A. W., "Electron Optics and the Electron Microscope," New York, John Wiley & Sons, 1945.
5. Boersch, H., *Naturwissenschaften*, **27**, 418 (1939).
6. Boersch, H., *Z. Tech. Phys.*, **20**, 346-350 (1939).
7. Zworykin, V. K., Hillier, J., and Snyder, R. L., *Am. Soc. Testing Mater. Bull.*, **117**, 15-23 (1942).
8. Hottenroth, G., *Ann. Physik.* **30** (5), 689-711 (1937).
9. Mahl, H., and Pendzich, A., *Z. Tech. Phys.*, **24**, 38-42 (1943).
10. Johansson, H., and Scherzer, O., *Z. Physik*, **80**, 183-202 (1933).
11. Mahl, H., *Jahrb. AEG Forsch.*, **7**, 43-56 (1940).
12. Freundlich, M. M., "Origin of the Electron Microscope," *Science*, **142**, #3589, 185-188 (1963).
13. Marton, C., Sass, S., Swerdlow, M., Van Bronckhorst, A., and Meryman, H., "Bibliography of Electron Microscopy," N.B.S. Circular 502, Aug. 1, 1950.

Cross-References: ELECTRON OPTICS, FIELD EMISSION, MICROSCOPE.

ELECTRON OPTICS

The invention of wave mechanics in 1926 by Heisenberg and Schrödinger saw a revolution in physics. It became apparent that the ideas of classical dynamics could be formally replaced when a stream of particles was considered. According to the well-known De Broglie hypothesis, a wavelength λ can be assigned to any material particle such that

$$\lambda = h/mv \quad (1)$$

where h is Planck's constant, m is the particle mass, and v is the particle velocity.

As one of the many consequences of these ideas, the new science of electron optics emerged.

In the same year, Busch demonstrated that the action of a short axially symmetrical magnetic field on a beam of electrons was similar to that of a glass lens on light. Terms used before then in optics found their use in describing electron devices. Electron "lenses" and "mirrors" having "focal lengths" and "resolutions and aberrations" were described. This analogy has proved useful in the study of the behavior of electrons in electronic valves, magnetrons and klystrons, traveling wave tubes, cathode ray tubes and electron microscopes, to name only a few devices. The original concept of the electron microscope was evolved by direct analogy with the light microscope.

An electron which has fallen through a potential V has a kinetic energy $\frac{1}{2}mv^2 = eV$. Hence $v = \left(\frac{2eV}{m}\right)^{1/2}$ and using Eq. (1), we obtain the useful formula expressing electron wavelength in terms of volts

$$\lambda = \left(\frac{150}{V}\right)^{1/2} \quad (2)$$

where λ is in angstroms.

In light optics, the least resolvable distance S between two objects is given by the Abbé expression $S = \frac{\lambda}{2n \sin i}$, where n is the refractive index of the material between object and lens and $2i$ is the angle subtended at the lens by the object; $n \sin i$ is called the numerical aperture. In the case of white light of equivalent wavelength $\lambda = 5600 \text{ Å}$,

$$S_{\text{(minimum, light)}} = 1800 \text{ Å, i.e., } 1.8 \times 10^{-5} \text{ cm}$$

Using the same expression and Eq. (2), which gives λ for an electron 0.04 Å for $V = 100 \text{ kV}$,

$$S_{\text{(minimum, electrons)}} \approx 0.04 \text{ Å}$$

i.e., small fractions of an angstrom can be resolved. However the theory does not take into account the fact that lenses are imperfect. Spherical aberration and chromatic aberration drastically affect the situation. In addition, certain diffraction defects and scattering impose limitations. These facts and others make it impossible to obtain resolutions near the theoretical maximum. Instead, 2.0 Å is a better theoretical value. Practical limiting resolutions obtained are of the order of 2.5 Å to 3.0 Å .

The basic source of electrons in electronic devices is the heated filament or disc. The well-known equation $J = AT^2 \exp\left(-\frac{\phi}{kT}\right)$ describes the emission, where J is the current density, T is the temperature, ϕ is the work function of the emitting material, A is a constant and k is Boltzmann's constant. J is enhanced by increasing T or reducing ϕ . Typical cathode materials used are tungsten, tantalum, the oxides of barium and strontium, and in certain cases, lanthanum hexabride. A voltage is applied to accelerate the

emitted electrons, and the current density is given by $J = AT^2 \exp\left(-\frac{\phi}{kT}\right)$ if they are all removed to the anode. However, many electrons stay near the cathode and repel other emitted electrons back to the cathode. Most devices operate in this "space-charge limited" way. For any electrode configuration, the current density is then $J = GV$ where V is the applied voltage and G is a constant for the particular electrode configuration. G is roughly equivalent to electrical conductance and is called the "perveance." The constant is fundamentally important. The higher G is in value, the greater is the efficiency of the beam system.

In simple designs, the axial portion of the cathode is overloaded producing excessive emission and cathode burnout. To avoid this situation, carefully shaped electrodes must be used, but the acceptable design depends on the application. A bent hairpin point cathode is used in, for example, the electron microscope. It produces low perveance but high emission density. Since G is low, the field at the cathode can be high, and this tends to reduce ϕ , giving emission at lower temperature. The electron guns for klystrons, traveling wave tubes, and metallurgical applications such as vacuum melting and welding require higher efficiency and current density. Hence they are high-perveance guns.

Electrostatic and magnetic fields control the motion of an electron according to the following equation:

$$\mathbf{F} = m \frac{d\mathbf{v}}{dt} = e[\mathbf{E} + \mathbf{v} \times \mathbf{B}] \quad (3)$$

where \mathbf{F} = force

\mathbf{E} = electrostatic field

\mathbf{B} = magnetic field

\mathbf{v} = electron velocity

e = electronic charge.

The electron must travel in a vacuum if it is not to be scattered and lose its kinetic energy by collision with relatively massive gas molecules. Nuclear particles such as protons and neutrons are each about 1837 times larger in mass than the electron.

Electrostatic Electron Lenses. If in Eq. (3),

$\mathbf{B} = 0$, then $\mathbf{F} = e\mathbf{E}$. The equation says that an electron in a field \mathbf{E} experiences a force \mathbf{F} in the direction of the field. Thus in the system shown in Fig. 1 where there is a voltage V_1 on cylinder 1 and V_2 on cylinder 2 and where $V_2 > V_1$, the field and the path of an electron are as shown. The electron moving from left to right increases its velocity and is deflected towards the axis. After passing the median, the force is away from the axis, but since velocity is increased, the electron spends less time in this part of the field. Therefore, deflection away from the axis is less than it was toward it, and there is a net convergence. If $V_2 < V_1$, as in Fig. 1(b), then for an electron beam traveling from left to right the lens will still be convergent because maximum deflection will occur after the electrons have slowed down. All such lenses are convergent for $E = 0$ on both sides of the lens.

Magnetic Electron Lenses. If in Eq. (3), $\mathbf{E} = 0$ then $\mathbf{F} = e\mathbf{v} \times \mathbf{B}$. This says that the force $\mathbf{F} = e|\mathbf{v}||\mathbf{B}|\sin\theta\mathbf{\hat{e}}$ where $|\mathbf{v}|$ and $|\mathbf{B}|$ are the numerical values of the vectors and θ is the angle between them; $\mathbf{\hat{e}}$ is a unit vector perpendicular to both, and indicates the direction of the force. For \mathbf{F} to be greater than 0, \mathbf{v} must be greater than 0. The force constrains the electron to move in a circle of radius $\rho = \frac{v \sin\theta}{B e/m}$. At the

same time, it moves in a perpendicular direction with velocity $v \cos\theta$ and, therefore, it traces the path of a helix and returns to the axis in a time $T = \frac{2\pi}{B e/m}$ which is independent of both \mathbf{v} and θ .

The net effect is that all electrons are focused by \mathbf{B} to produce an image of the source from which they diverge. Figures 2(a) and 2(b) illustrate a simple magnetic lens which consists of a short coil of wire contained in a surrounding shield of magnetic material. A short gap in the magnetic material concentrates the escaping field when the electromagnet is energized. A spiral distortion of the electron beam results from the use of this sort of lens. For example, in an electron microscope, dual magnetic lenses [Fig. 2(c)] are therefore used. The coils are wound in opposite sense to make distortions cancel. Figure 3 illustrates an electron microscope.

A device worthy of brief consideration is the cathode ray tube. This is used to display voltage wave forms. Its basic components are as follows:

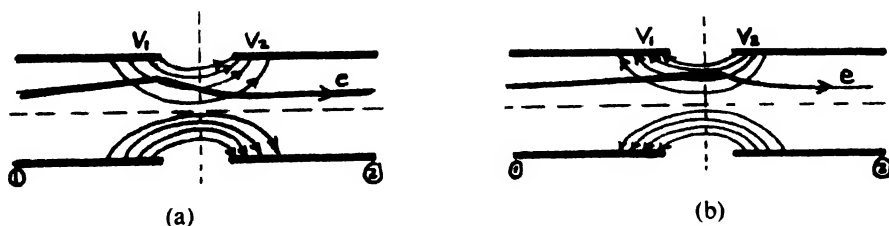


FIG. 1a and b. Electrostatic focusing of electron beam by cylinder lenses (1) and (2).
Fig. 1a, voltage $V_2 > V_1$; Fig. 1b, voltage $V_2 < V_1$.

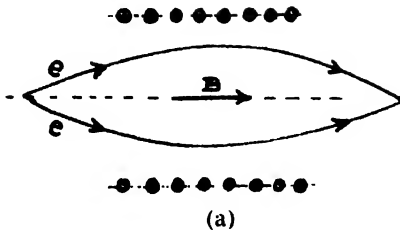


FIG. 2a. Magnetic focusing of electron beam by solenoid field. Electron path is a spiral as it moves from left to right. Velocity along axis is $V \cos \theta$.

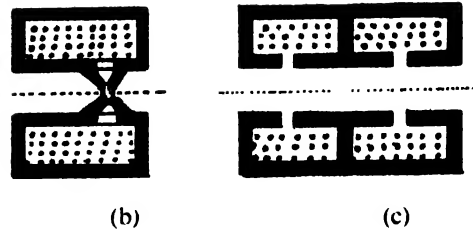


FIG. 2b and c. Magnetic electron lenses. The magnetic field produced by the coil windings is mostly contained in the surrounding shield of magnetic material. The useful volume of the lens is where the field jumps the gap in the shield.

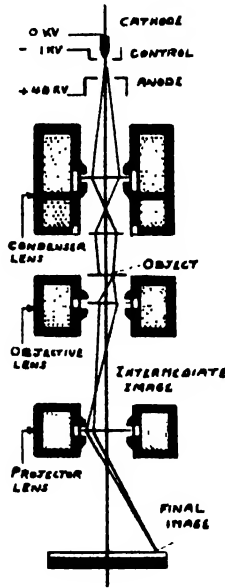


FIG. 3. Schematic diagram illustrating the general arrangement of a transmission electron microscope.

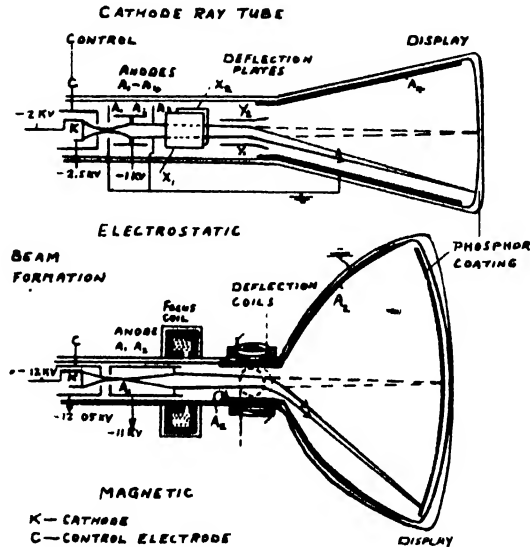


FIG. 4a. Schematic diagram of cathode ray tubes.

an electron gun, an acceleration and focus system, a deflection, and a display system. Figure 4(a) shows the construction. Depending on the type of tube, the applied voltage generates an electrostatic or magnetic field, and the field deflects a beam of electrons on its path to a fluorescent screen. The deflection is proportional to the voltage. In the television tube, the beam is deflected to trace each part of the screen in sequence. At the same time it is intensity modulated, enabling the picture to be constructed of dark and light elements. Magnetic deflection is used in this case because large scan angles are required.

Electrostatic Deflection. If in Fig. 4(b) a voltage V_d exists between the plates x_1x_2 and an electron enters along the axis with a constant velocity $v = (2Ve/m)^{1/2}$, then because of the field E_y caused by V_d , a transverse force F_y causes a

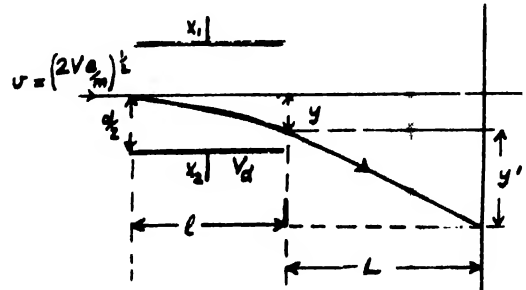


FIG. 4b. Deflection of electron beam by electric field between parallel plates X_1X_2 .

motion in this direction to be impinged on the electron, i.e., $m \frac{d^2y}{dt^2} = \frac{V_a e}{d}$, from which $\frac{dy}{dt} = \frac{V_a e}{d} t$ and $y = \frac{1}{2} \frac{V_a e}{d} t^2$. The force acts for time $t = \frac{l}{v}$; thus deflection $y = \frac{1}{2} \frac{V_a e}{d} m \left(\frac{l}{v} \right)^2 = \frac{1}{4} \frac{V_a l^2}{d} \frac{1}{v}$. The additional deflection, after F has ceased to act on the electron, is $y' = \frac{V_a l L}{V} \frac{1}{2d}$.

Magnetic Deflection. See Fig. 4(c). If B is a sharply defined field, then the deflection on leaving the field is given by $y = \frac{l^2}{2\rho}$. Total deflection $= D - L \tan \alpha + y$. To a good approximation $D = \frac{BLl}{v} \frac{e}{m} \left(1 + \frac{l}{2L} \right)$.

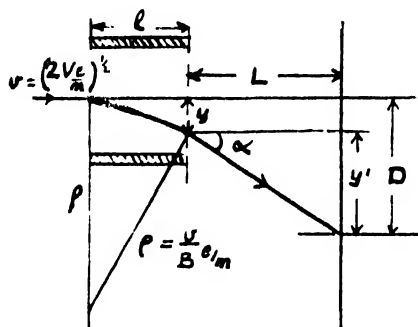


FIG. 4c. Deflection of electron beam by magnetic field. Electron path is an arc of a circle only while it is influenced by field. After leaving the field the path is linear.

It should be emphasized that these considerations are only approximate. In practice, factors such as the inability to sharply define a field edge in space cause added complications.

BARRY A. GEORGE

References

1. Klemperer, "Electron Optics," Cambridge, The University Press, 1953 (for a rigorous treatment of electron optics).
2. Bakish, R., "Introduction to Electron Beam Technology," New York, John Wiley & Sons, 1962. (for the practical applications of electron optics).
3. Bakish, R., Ed., "Electron and Ion Beams in Science and Technology," New York, John Wiley & Sons, 1965.
4. Pierce, J. R., "Theory and Design of Electron Beams, Princeton, N. J., D. Van Nostrand Co., 1949.
5. Spangenberg, K. R., "Vacuum Tubes," New York, McGraw Hill, 1948.

Cross-references: ELECTRON; ELECTRON MICROSCOPE; ELECTRON TUBES, RECEIVING TYPE; OPTICS, GEOMETRICAL; OSCILLOSCOPE; THERMIONICS.

ELECTRON SPIN

The electron is an elementary particle of essential importance to atomic physics and, in addition to its stability, basic charge e , and small rest mass m , it has been found to possess a quantum mechanical attribute named "spin." As the name implies, this is a mechanical angular momentum of fixed magnitude (projected) of $\frac{1}{2}h$ (one half Planck's constant divided by $2\pi = 0.52722 \cdot 10^{-31}$ joule-second). Intimately associated with this intrinsic quantized angular momentum is a magnetic moment of approximate value $eh/2m$ (this quantity is called the *Bohr magneton*, μ_B , and has the value $0.92732 \cdot 10^{-23}$, ampere-meter²). While these concepts have a classical analogy in an imagined spinning or rotational motion of the charge and mass of the electron (assumed to be not a point), nevertheless, it is not possible to treat the electron spin wholly classically. For instance, the electron *gyromagnetic ratio* (ratio of the magnetic moment to the mechanical moment) is twice the classical value. Also, since the spin quantum number is limited to $\frac{1}{2}$, the spin disappears in the classical limit of $h \rightarrow 0$. The fact that these properties are quantized implies their quantum nature, and indeed, the spin and magnetic moment are natural consequences of the relativistically invariant Dirac equation.

The hypothesis of the electron spin was first proposed in 1925 by Uhlenbeck and Goudsmit to explain the spectroscopic fine structure. The spin can be introduced as a vector S such that the square of the spin has the eigenvalue $S \cdot S = s(s+1)h^2$ where $s = \frac{1}{2}$ so that $S = \sqrt{\frac{3}{4}}h$. The component of S in a specified direction, such as the z axis, is $S_z = mh$ where $m_s = +\frac{1}{2}$ or $-\frac{1}{2}$. This implies a spin angular momentum component of absolute value $\frac{1}{2}h$. It is further necessary to postulate that the intrinsic magnetic moment of the electron (due to its spin) is $\mu = -e/mS$. The component of μ along the z axis is then $\mu_z = m_s(eh/m)$ or $\pm \mu_B = \mu_B$.

Direct experimental evidence for quantized electron angular momentum and magnetic moment was presented by the Stern-Gerlach experiment in 1922. A collimated beam of neutral silver atoms was directed through an inhomogeneous magnetic field. This field acted on the magnetic moment of the unpaired valence electron of the silver atom to produce a transverse force and, hence, a lateral deflection. In contrast to the classical expectation of a continuous spread of deflections due to many possible orientations of the electron magnetic moment, it was observed that only two opposite deflections occurred. This implied only two possible values for μ_z leading to the space quantization of spin, $m_s = \pm \frac{1}{2}$. Similar molecular beams have provided Rabi and Kusch and their colleagues with a

supply of free atoms and molecules whose orientation in magnetic fields could be varied with magnetic resonance. Considerable information on atomic hyperfine structure has resulted from these measurements as well as recognition of the anomalous magnetic moment of the electron.

It had been thought that the magnetic moment of the *free* electron could not be directly measured due to the overwhelming force associated with the motion of the charge in the magnetic field. However, Crane and his colleagues have trapped electrons into helical orbits in a magnetic field and have accurately compared the cyclotron orbital frequency to the spin precessional frequency. This makes possible a highly precise measurement of the *electron magnetic moment* and its value was found to be slightly larger than a Bohr magneton, $\mu_e = 1.001159622 \mu_B$. This anomaly is now understood as a consequence of the uncertainty jittering or pulsation in which the electron exchanges virtual photons with the radiation field and experiences a sort of radiation reaction which increases its effective inertial mass. The spinning motion should not be affected by this change and, hence, the magnetic moment should be slightly greater when measured in units of Bohr magnetons using the apparent inertial mass of the electrons. The theory has been worked out in adequate detail and the electron magnetic moment is predicted to be $\mu_e = \mu_B(1 + \alpha/2\pi - 0.327\alpha^2/\pi^2)$ in good agreement with experiment. This more precise value is 0.92839×10^{-23} ampere-meter².

An important consequence of the spin concept occurs when the Pauli exclusion principle is applied to polyelectronic atoms using the spin quantum number $m_s = \pm \frac{1}{2}$ in addition to the principle quantum number n , the azimuthal quantum number $l \leq n - 1$, and its projection $m_l \leq l$. The Pauli exclusion principle; based on the symmetry properties of the wave functions appropriate to the electron as a Fermi particle (i.e., spin $\frac{1}{2}$ particle), limits the number of electrons in an atom to those which can be differentiated by unique sets of quantum numbers. The possible combinations of the integral quantum numbers n, l, m_l with the two possible values of m_s determine the number and structure of the electrons in the atom. The recurring similarities in the orbital characteristics of the outer electrons produced by the quantum numbers determine the valencies and chemical characteristics of the atoms in a periodic fashion as exemplified in the periodic table.

Since the electron spin can have two orientations with respect to the axis of the orbital motion of the electron, a difference in energy will result from the interaction of the spin magnetic moment with the magnetic field produced by the orbital motion. This energy difference can be evaluated and leads to a *fine structure separation energy* which is less than the electronic level separations by a factor of the order of α^2 where $\alpha = e^2/4\pi\epsilon_0\hbar c$ is the fine structure constant, approximately $1/137$. This is the origin of the well-known sodium yellow doublet.

The imposition of an external magnetic field gives rise to further splittings called *Zeeman effects*. The spin couples with the orbital motion giving rise to *splitting factors*, g , which are complicated because of the different gyromagnetic ratios of the spin and orbital motions. If the magnetic field is strong enough, it can uncouple the spin from the orbital motion giving a somewhat simpler Zeeman splitting called the complete Paschen-Back effect.

The magnetic effects of the electron's spin magnetic moment as well as their orbital moment are reduced by the tendency of electrons in atoms to pair spins of opposite direction and to complete shells of compensating orbital motion. This is even more pronounced in molecules since the chemical binding force is essentially the result of pairing electrons and completing shells. Nevertheless, when an unpaired electron exists, its magnetic moment is available for detection and can be measured by determining the aggregate magnetic moment of the sample. However, a more sensitive method, with considerably better accuracy and resolution, involves a magnetic resonance measurement (see MAGNETIC RESONANCE). When applied to the electron spin magnetic moment the measurement is often called ESR, or *electron spin resonance*, but more widely, EPR, or *electron paramagnetic resonance*. These techniques extend the earlier atomic and molecular beam resonance measurements of Rabi and his school.

The detection can be described, quantum mechanically, in terms of the absorption of photons of frequency f and photon quantum energy hf by excitation of transitions between neighboring quantum states of the atom or molecule separated by the Zeeman splitting due to an imposed external magnetic field. Since the energy difference is primarily $g\mu_B B = hf$, the frequency for reasonable fields of the order of a thousand gauss (i.e., $B \approx 0.1$ weber/m²) is in the microwave region, i.e., $f \approx 10^{10}$ cps. Consequently, the techniques appropriate to these measurements are associated with microwave spectroscopy.

The simplest cases are those in which the odd electrons are in *S* states with no orbital motion, or in which the magnetic effect of the orbital motion is "quenched" by the presence of neighboring atoms. Then, "spin-only" effects occur and the splitting factor g is close to 2. When quenching does not occur, the orbital motion magnetic field complicates the Zeeman splittings and gives rise to more complex patterns of resonance. In solids, the crystalline electric field affects the orbital motion and, hence, the pattern which then varies with crystal orientation. Interaction of the electron magnetic moments with the nuclear magnetic moments gives hyperfine splittings with even greater detail.

A large number of atoms which show paramagnetic or electron spin effects are those in the transition groups, i.e., iron group, rare earth, palladium group, platinum group and actinide group in which inner electron shells are being filled.

Other situations where the electron magnetic moment effects are useful for analysis or detection include free radicals, molecules with broken bands from irradiation or heating, etc., conduction electrons in some metals, and defects and impurities in crystals.

In solids composed of elements of the iron group (occasionally the rare earths), the long-range cooperative effects between electron spin magnetic moments give rise to ferro-, ferri-, and antiferromagnetism.

WILLIAM E. STEPHENS

References

Electron Spin

- Harnwell, G. P., and Stephens, W. E., "Atomic Physics," p. 74, New York, McGraw-Hill Book Co., 1955.
- Leighton, R. B., "Principles of Modern Physics," pp. 89, 184, 667, New York, McGraw-Hill Book Co., 1959.
- Van Vleck, J. H., "Electric and Magnetic Susceptibilities," Oxford, Clarendon Press, 1932.
- Ramsey, N. E., "Molecular Beams," Oxford, Clarendon Press, 1956.
- Uhlenbeck, G. H., and Goudsmit, S. A., *Physica*, **5**, 266 (1925); *Nature* **117**, 264 (1926).
- Stern, O., and Gerlach, W., *Ann. Physik*, **74**, 673 (1924); *Z. Physik*, **41**, 563 (1927).
- Wilkinson, D. T., and Crane, H. R., *Phys. Rev.*, **130**, 852 (1963).

Electron Spin Resonance

- Pake, G. E., "Paramagnetic Resonance," New York, W. A. Benjamin, Inc., 1962.
- Low, W., "Paramagnetic Resonance in Solids," New York, Academic Press, 1960.
- Bleaney, B., and Stevens, K. W. H., *Rept. Progr. Phys.*, **16**, 108 (1953).
- Bowers, K. D., and Owen, J., *Rept. Progr. Phys.*, **18**, 304 (1955).
- Orton, J. W., *Rept. Progr. Phys.*, **22**, 204 (1959).

Cross-references: ELECTRON, MAGNETIC RESONANCE, ZEEMAN AND STARK EFFECTS.

ELECTRON TUBES, RECEIVING TYPE

An electron tube consists of a heater for kinetic energy excitation of electrons, a cathode which acts as a transfer electrode source of electrons, controlling grid electrodes, and an anode that is maintained electrically positive with respect to the cathode. These elements are insulated from each other and enclosed within an evacuated envelope made of either glass, metal, ceramic, or a combination of these materials. A getter is flashed within the tube to absorb any residual gas molecules which could have a harmful effect—electrically and chemically—on the operation of the tube.

When the device has only two electrodes, a

cathode and an anode, it is called a *diode*. With the anode maintained electrically positive with respect to the cathode, an electric field results which causes the electrons to move toward the anode. In the external circuit, the electrons flow from the anode through the load impedance and then through the voltage source to the cathode, which acts as a low-work-function transfer medium, and so back to the anode. In this discussion, the work function is considered to be the total amount of work necessary to free an electron from a solid.

Other electrodes are often introduced between the cathode and the anode in the form of grids. By varying the voltages on these intervening electrodes, it is possible to modify the electric field between the cathode and the anode, and thus to control the current in the external circuit. Tubes having one grid in addition to the cathode and anode are called *triodes*. Tubes with two grids are called *tetrodes*; tubes with three grids are called *pentodes*. In general, tubes are labeled in accordance with the total number of active electrodes in a linear arrangement using a common electron stream. Sometimes, two or more sections are enclosed within the same envelope (e.g., a diode-triode or a triode-pentode); these tubes are not referred to in terms of the total multi-electrode structure, but they are designated in terms of the respective tube units.

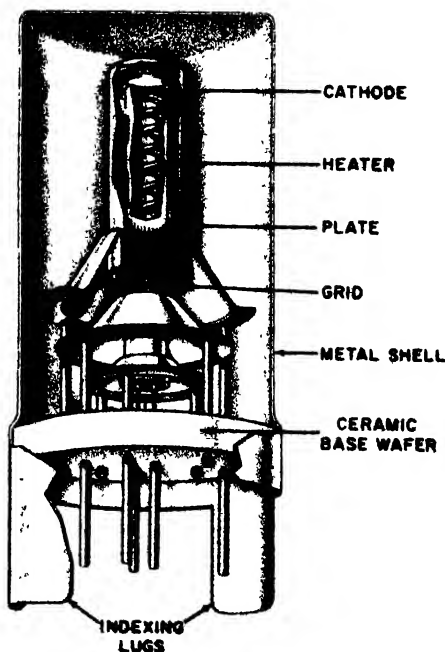


FIG. 1.

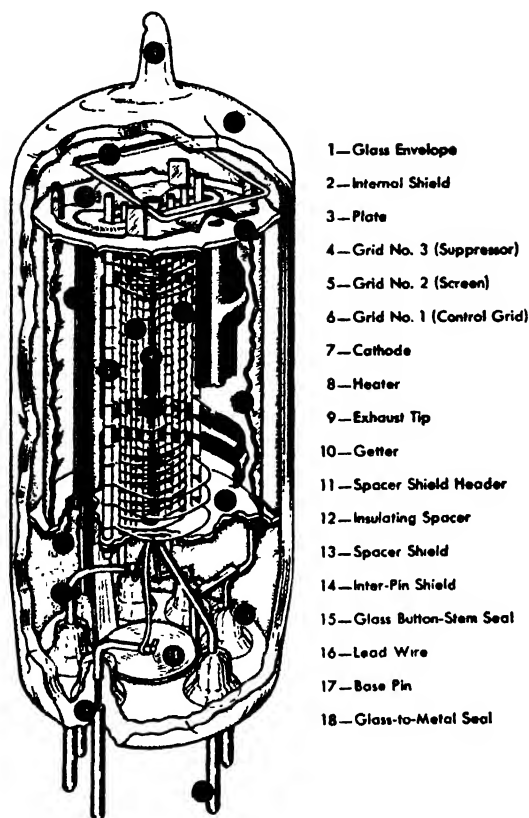


FIG. 2. Structure of a miniature tube.

Oxide-coated Cathode. Oxide-coated filamentary cathodes can operate at relatively low temperatures of 1000°K because of the low-work-function surface layer. However, they are subject to sputter effects and can also evaporate substantial amounts of material that deposit onto adjoining electrodes and lead to harmful grid-emission and contact-potential phenomena. Such oxide-coated filaments are high-efficiency emitters at low-wattage inputs and are used successfully in low-current pulsed high-voltage rectifiers for scanning systems in television receivers.

The indirectly heated cathode consists of a nickel alloy sleeve coated with alkali earth oxides of barium and strontium, and, inside the sleeve, a heater of alumina-coated tungsten or molybdenum-tungsten alloy wire. The heater wire is in the form of a helical coil or folded strands; it is coated with alumina to insulate the heater wire from the cathode nickel sleeve. In addition this insulating coating prevents adjoining helix turns or strands of wire from short-circuiting each other. The cathode sleeve is heated by conduction

and radiation from the heater. Because the oxide-coated cathode is electrically isolated from the heater, it is called a *unipotential cathode*, since unlike the filamentary type, there is no voltage drop along its length due to heater current.

The use of the indirectly heated unipotential cathode offers two advantages. The nickel-alloy sleeve acts as a magnetic shield around the heater wire and minimizes the effects of 60-cycle hum from the ac heater supply. Secondly, the use of separate heater and cathode circuits permits the design of close-spaced control grids and cathodes which in turn result in higher gain amplifier service. In addition, the use of close cathode-to-anode spacing design in rectifier service creates low tube-voltage drops and better voltage regulation.

Because of these advantages, electron tube manufacturers make extensive use of the indirectly heated cathode system. The oxide coating formed on the nickel-alloy cathode is derived from alkaline earth carbonates, either 50/50 mole per cent barium-strontium carbonate or 49/45/6 mole per cent barium-strontium-calcium carbonate. The presence of this activated oxide coating lowers the work function of the substrate nickel.

The electron emission performance of the indirectly heated cathode is influenced by three basic factors: (1) the effects of the low work function, (2) an *n*-type semiconductor method of electron transfer, and (3) the electron transfer efficiency involved in the porous nature of the alkaline earth oxide matrix coating. These three factors are discussed in the following paragraphs.

The work function ϕ of a material is measured in electron volts and is temperature dependent. Such temperature dependence can be influenced by the work function of the substrate metal or by adsorbed layers of atomic films. When atoms having a lower work function are adsorbed onto the surface of a metal having a higher work function, the work function of the substrate metal is reduced by the decrease in its surface energy potential barrier. For example, when a monolayer of barium ($\phi = 2.52 \text{ eV}$) is adsorbed on the surface of nickel ($\phi = 4.96 \text{ eV}$), the resultant adsorption layer of barium-oxygen-nickel has a work function $\phi = 0.9 \text{ eV}$.

The usefulness of materials having low work functions is limited by their high rate of evaporation which tends to deposit material onto adjoining electrodes as well as to shorten the life of the system under vacuum-tube conditions. Accordingly, balance of work function and rate of evaporation is a desired feature in effective cathode design. The barium/barium oxide component of the alkaline earth oxide matrix meets this requirement by having the lowest usable work function consistent with a minimum rate of evaporation and long life at the operating temperature (1025°K) of such cathode systems.

Normally, barium oxide is an insulator. In vacuum-tube technology, however, the oxide is made to act as a semiconductor material by an activating process which uses physical chemical reactions to form barium and associated donor

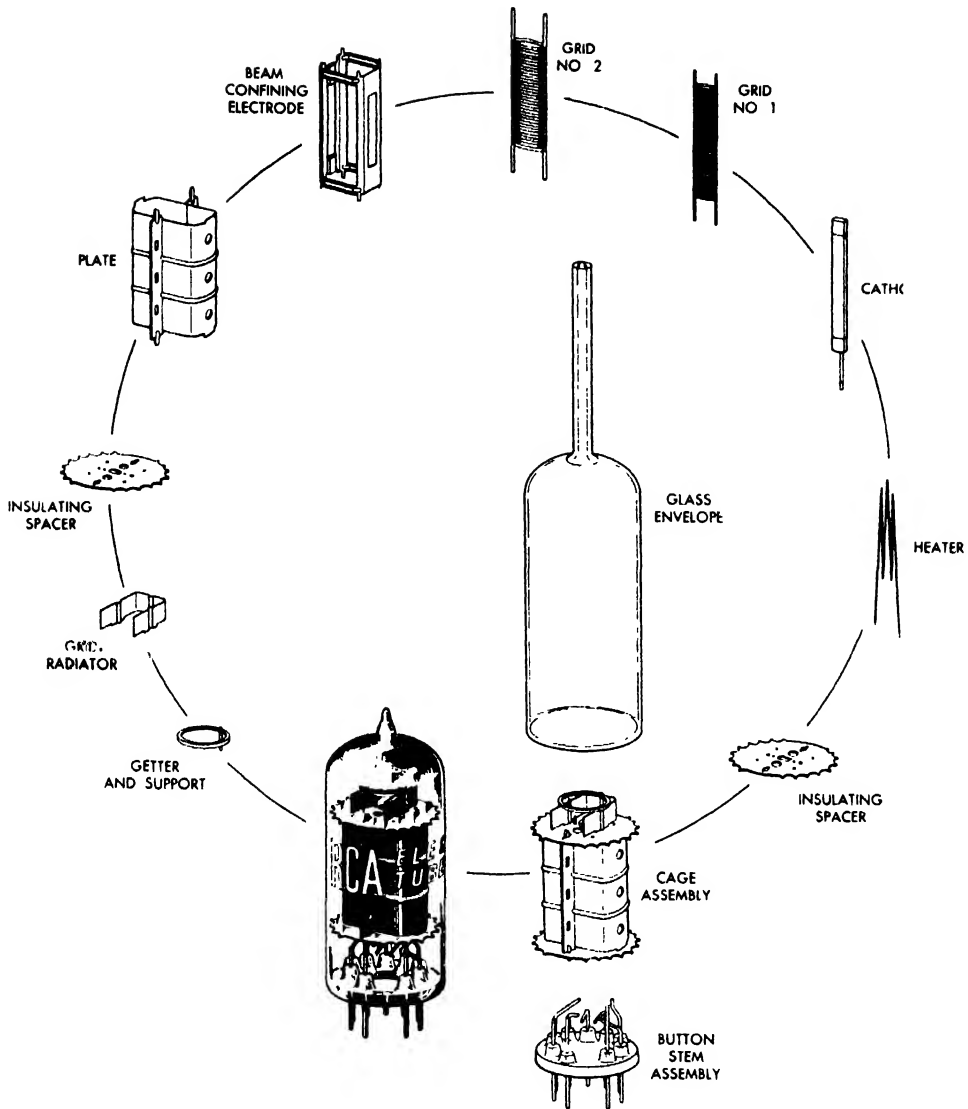


FIG. 3. Parts of a Novar Tube.

sites in the oxide lattice. In addition to having a low work function, the barium oxide now exhibits properties similar to an *n*-type semiconductor which allows it to transfer electrons at temperatures below 575°K. Furthermore, a relatively thick oxide coating of 0.5 to 2.0 mil thickness will result in favorable coating porosity and electron pore gas transfer at temperatures above 575°K. Although electron pore gas transfer does not reduce the work function of the system, it does improve the conductivity of the oxide layer at 1025°K. The resulting decrease in voltage drop across the oxide coating reduces the potential minimum and improves the emission performance.

Barium and the associated donor centers are

produced by the chemical reduction of barium oxide by elemental agents of carbon, magnesium, silicon, and tungsten present within the cathode nickel alloy. This reproductive action is a function of the effective concentration of the elements, as well as their rate of diffusion through the metal, and their rate of reaction with the emission oxides at the interface region.

In conclusion, the operating characteristics of the oxide cathode system itself are determined by a number of physical chemical factors including: the porosity of the applied coating; the conversion of the carbonates to the oxide crystal lattice form; the minimum sintering action induced by the eutectic phase of barium carbonate/barium oxide; the rate of electrolytic transport

of the barium ions; the evolution of oxygen gas; the rate of evaporation of barium/barium oxide; and the formation of films on adjoining electrodes. Thus, the oxide-coated cathode system operates in a dynamic equilibrium involving solid, ion, and gas phase changes across the interface region of the cathode metal-to-oxide layer as well as the phase boundary between the oxide layer and the residual gases and vapors in the vacuum regions of the tube. These equilibria are dependent upon the noninjurious trace concentrations of carbon dioxide, oxygen, water vapor, sulfur, halogens, and volatile metal oxides in the environmental vacuum region for the best electron transfer performance and long life of the cathode system.

Diodes. As previously mentioned, the diode is the simplest form of vacuum tube having a cathode system and anode together with a flashed getter. In well-degassed tube structures, the reducing element content of the cathode nickel alloy creates sufficient active barium sites in the emission oxide matrix so that the unit itself can function as a getter at high operating temperatures. Diodes are used as rectifiers, detectors, dampers, and limiters. These tubes are high-vacuum types in which the internal voltage drop is proportional to the dc load current. When a low constant-voltage drop is desired, mercury vapor tubes are used. The constant-voltage drop of 15 volts is a function of the ionization potential of the mercury vapor since the positively charged mercury ions neutralize the space charge effect. In general, for diodes, when the current demand is less than the temperature-limited value, the current I will vary as the three-halves power of the applied voltage E as given by the following equations. For the plane parallel system: $I = (2.33 \times 10^{-6} E^{3/2}) S^2$ where S is the cathode-to-anode spacing. For the concentric cylindrical system: $I = (14.65 \times 10^{-6} E^{3/2}) (h^2 \times r_a)$ where h is the ratio of the anode radius r_a to the cathode radius r_c . For very low anode voltages, the above approximations are not accurate because the effects of the initial electron velocity and the contact potential result in currents larger than calculated.

Not all electrons leaving the cathode reach the anode. Some electrons return to the cathode by reason of a balance between the initial electron velocity and the image force; i.e., when an electron is emitted from a metal surface and is at some distance from the surface, it induces a charge of equal magnitude but opposite sign in the interior of the metal. As a result, some electrons remain in the space above the cathode and produce a cloud of electrons or "space charge" which repels some electrons back to the cathode and thus impedes the flow of electrons to the anode. The extent of this action and the amount of the space charge are dependent upon the temperature of the cathode, the cathode-to-anode spacing, and the anode potential.

Under fixed temperature conditions, the maximum number of electrons that are emitted is also fixed. The higher the anode potential, how-

ever, the lower is the number of electrons remaining in the space charge region; as a result, an increase of the anode voltage results in an increase in current until saturation is reached. Beyond this condition (maximum current at maximum voltage at fixed cathode temperature conditions), additional anode (plate) voltage will only increase the plate current slightly because of the reduction of the work function at the cathode due to the electrostatic effect of the applied field.

The Richardson-Dushman equation yields information pertaining to the work function ϕ of cathode systems and the zero field saturated current I_{s0} with respect to temperature T . The emission equation for a cathode of surface area

S has the form $I_{s0} = A_0 S T^2 e^{-\frac{11610 \phi_0}{T}}$. The zero field saturated current I_{s0} is extrapolated from a plot of the log I_s (saturated current) as a function of the square root of the applied voltage at the anode V_a . The work function with an external field E is a function of the square root of the external field strength such that $\phi_E = \phi_0 - 3.78 \times 10^{-5} E^{1/2}$.

Triode. When a third electrode, called a grid, is placed between the cathode and the anode, the tube is called a triode. This grid consists of a fine wire wound on two support side rods; the spacing between the turns of the wire is relatively large so that electrons are able to pass from the cathode to the anode. When the tube is used as an amplifier, a varying signal imposed on a negative dc voltage on the grid controls the plate current. As the grid voltage becomes more negative, the plate current decreases while the plate voltage increases back to the original applied potential as a result of the decrease in the voltage drop across the load impedance. In other words, the grid voltage and the plate current are in phase, but the plate voltage is 180° out of phase with the grid voltage. The cathode, grid, and plate of the triode form an *electrostatic system* such that it is possible to equate the plate current I_b to a three-halves power law similar to that for diodes, i.e.,

$$I_b = K \left[\frac{C_{gk}}{C_{pk}} \cdot E_g + E_b \right]^{3/2}$$

where K is a constant (the permeance) which depends upon the geometry of the tube, and C_{gk} and C_{pk} are the grid-to-cathode and the plate-to-cathode capacitance, respectively. This equation assumes a negligible effect for the initial velocity of electrons from the cathode. Actually, the initial electron velocity creates the space-charge effect so that the cathode is slightly negative (instead of zero) and a potential minimum exists a short distance in front of the cathode. The effect of this shift is to increase the plate current slightly. Because the initial electron velocity distribution depends upon cathode temperature, the plate current will also be influenced by the cathode temperature even though it is limited by the space charge.

The electrical characteristics of triodes are described in terms of three parameters: the amplification factor (μ), the dynamic plate resistance (r_p), and the transconductance (g_m) as given by the following relationships:

$$\mu = \frac{C_{pk}}{C_{pk}} = \frac{\partial e_b}{\partial e_g} = - \frac{de_b}{de_g} \Big|_{i_b \text{ constant}}$$

$$r_p = \frac{\partial e_b}{\partial i_b} = \frac{de_b}{di_b} \Big|_{e_g \text{ constant}}$$

$$g_m = \frac{\partial i_b}{\partial e_b} = \frac{di_b}{de_b} \Big|_{e_g \text{ constant}}$$

These relationships are applicable within the region where the family of curves is straight, parallel, and equidistant for equal increments in the parameter. The triode tube can be considered a linear circuit element within small variations, and the quantities μ , r_p , g_m may be used as constants in the analysis of tube performance. A useful approximation for the calculation of plate current I_b is the expression

$$I_b = \frac{2.33 \cdot 10^{-6} (E_g + E_b/\mu)^{3/2}}{(S_{rk})^2} \quad \text{or} \quad I_b = \frac{2.33 \cdot 10^{-6} \left(E_g + \frac{E_b}{\mu} \right)^{3/2}}{(S_{rp})^2/\mu}$$

where S is the distance between grid-to-cathode and grid-to-plate, respectively and

$$\mu = \frac{g_m E_g}{3/2 I_b - g_m E_g}$$

Tetrodes. The effects of interelectrode capacitances between the grid and the plate and between the grid and the cathode sometimes result in coupling between the input and output circuits. Such coupling and impedance mismatching can cause instability and low output performance. The grid-to-plate capacitance can be sufficiently reduced by the introduction of an additional grid electrode, called the screen grid, between the control grid and the plate, and such tubes are known as tetrodes. In practice, the control grid-to-plate capacitance is reduced from several picofarads for a triode to less than 0.01 pF for a screen grid tube when a rf capacitor is connected between the screen grid circuit and the cathode.

The screen grid operates at a positive voltage and supplies an electrostatic force which pulls electrons from the space charge region, through its widely spaced turns of wire, and onto the plate. The plate is shielded from the cathode and exerts insignificant force on the space charge region. The plate current in a tetrode is almost independent of the plate voltage and depends mainly upon the screen grid voltage. With tetrodes, moderately high amplification can be obtained without capacitive feedback from the plate to the control grid. However, for operation

in amplifiers where linearity is required between the control grid voltage and the plate current, the transconductance must be constant. To achieve this condition, the tetrode must be operated in the region where the plate voltage is higher than the screen voltage, i.e., where the family of curves is nearly straight, parallel, and equidistant for equal increments.

Tetrodes are not useful for amplification of very large signals because the proximity of the positive screen grid to the plate (at equal or slightly higher positive voltage) permits the capture of secondary electrons emitted from the plate. The capture of the lower-velocity secondary electrons by the screen grid is more pronounced when the plate voltage swings lower than the screen grid voltage. This condition occurs when large signal variations cause an increased voltage drop across the load impedance in the plate circuit.

Pentodes. The effects of secondary emission from the plate in a tetrode configuration are minimized when a fifth electrode is used. This fifth electrode (a third grid) is made of widely spaced turns of wire between the screen grid and the plate, and is known as the *suppressor grid*; it is usually connected to the cathode at zero potential. A tube with five electrodes in a linear arrangement, a cathode, three grids, and a plate, is called a pentode. The suppressor grid is negative with respect to the plate and can divert secondary electrons of low velocity back to the plate. Pentodes are capable of high voltage amplification and high power output at low levels of driving voltage on the control grid. When large signal inputs are placed on the control grid, large load impedance voltage drops result in the plate circuit which cause the plate voltage to become momentarily lower than the screen grid voltage without the loss of secondary electrons from the plate.

Instead of a suppressor grid, the interelectrode region between the screen grid and the plate can be virtually used as a space charge region to minimize the loss of secondary electrons from the plate. In practice, a suppressor grid or a beam-confining plate is used to enhance the effect of the electron cloud density. This effect is created primarily by designing the tube electrodes in such a way that the electron stream from the cathode is confined and concentrated in the region between the screen grid and the plate. The space potential is depressed at some point in the interelectrode region in front of the anode to a value below the anode potential to prevent secondary electrons from reaching the screen grid. A tube with this feature of construction is called a *beam-power tube*.

In beam-power tubes, the pitches of the helical turns of the lateral wire of the control grid and the screen grid are made equal and the wires are aligned to confine the passage of electrons into flat beams. A focusing beam plate, or optional use of a wide-spaced suppressor grid electrode, is used to maintain this high electron density. Because of the alignment of the control grid and

the screen grid, very few electrons are captured by the screen grid; thus, a beam power tube draws low current in the screen grid circuit. Effective suppressor grid action and low screen grid current permit beam power tubes of moderate size and voltages to operate efficiently at high power outputs with large values of transconductance. The large values of transconductances obtained with beam power tubes make them useful in the design of wide-band amplifiers for television service.

In both tetrodes and pentodes, the control grid is sometimes constructed in a non-uniform manner. The spacing between the adjoining helix turns of the grid lateral wire is made closer at the ends than at the center of the grid. In tubes with such variable pitch grids, the amplification factor decreases when the plate current decreases. Such grid structures also require large negative voltages to produce cut off conditions. These tube types are known as remote cutoff or variable- μ tubes. Tubes having uniform grids are called sharp cutoff tubes. Remote cutoff tubes are used in automatic-volume-control circuits in which the dc component of the control grid voltage is varied in such a manner that the amplification of the tube is made smaller for large signal voltages.

Multifunction Tubes. In addition to the single function multigrid tube such as a tetrode amplifier 6CY5 or a pentode amplifier 6BC5, there are other multi-electrode tubes that perform more than one function, e.g., an oscillator and mixer in a superheterodyne receiver. The term pentagrid converter represents a tube with five grid structures that lie in the electron stream between the cathode and the plate. Two of the grids function as control grids; the remaining grids function as screen or suppressor grids to shield the control grids from the plate, e.g., pentagrid converters 6BE6 and 6SA7 or the 6BY6, a pentagrid amplifier used in color television receivers as a sync separator and sync clipper as part of a gated amplifier circuit.

There are also multi-unit tubes consisting of combinations of diodes, triodes and pentodes in one envelope which use more than one cathode and/or plate. For example, the twin triode 12AX7A used in resistance coupled amplifiers; the 8AW8A in which the triode unit is used in a sync separator circuit and the pentode unit is used as an i.f. or video amplifier; and the triple triode 6EZ8 used in oscillator mixer and AFC circuits in FM receivers.

Microwave. The above mentioned multigrid and multi-unit tubes normally are used in amplification circuits at frequencies up to 500 Mc. However, above this frequency range, the transit time of electrons is short when compared to the voltage cycle and, as a result, the phase angle between the plate current and the plate voltage becomes less than 180° . Accordingly, an increase occurs in the power dissipated at the plate for a given designed power output. At these very high frequencies, electron loading can occur at the control grid because of the short transit time of the electrons from the cathode to the plate.

Certain pencil triode tubes are made to operate in the 500-Mc region with regular tube constructional arrangement except that the interelectrode spacing between cathode and grid, and between grid and plate, is made very small—about 1-mil clearance—in order to accommodate the short transit time of electrons.

To avoid the difficulties resulting from the short transit time of electrons as well as the effects of interelectrode capacitances, microwave tubes have been designed to make use of velocity-modulated beams of electrons. Klystrons, traveling-wave tubes, and magnetrons are special-purpose tubes designed to operate in the microwave region.

CARL H. MELTZER

References

- Spangenberg, K. R., "Vacuum Tubes," New York, McGraw-Hill Book Co., 1948.
- Deketh, J., "Fundamentals of Radio Valve Technique," Eindhoven, N.V. Philips, 1949.
- Loosjes, R., and Vink, H., "Properties of Pore Conductors," *Philips Res. Rep.*, 4, 449 (1949).
- Nergaard, L. S., "Oxide Cathode," *RCA Rev.*, 13, 464 (1952); "Physics of the Cathode," *RCA Rev.*, 18, 486 (1957); "Thermionic Emitters," *RCA Rev.*, 20, 191 (1959).
- Knoll, M., "Materials and Processes of Electron Devices," Berlin, Springer-Verlag, 1959.
- Langford-Smith, Ed., "Radiotron Designers Handbook," Fourth edition, RCA, 1960.

Cross-references: CIRCUITRY, DIODE (SEMICONDUCTOR), POTENTIAL, TRANSISTOR.

ELECTRONICS

The term "electronics" (Ger. *Elektronik*) was first used to describe the branch of physics now generally called physical electronics; that usage goes back almost to the discovery of the electron in 1897, as witness the names of two early journals in the field, *Jahrbuch der Radioaktivität und Elektronik* (founded in 1904) and *Ion: A Journal of Electronics, Atomistics, Ionology, Radioactivity and Raumchemistry* (1908).

In the currently prevalent technological context, the adjective "electronic" and the noun "electronics" (Ger. *technische Elektronik*) date back only to the 1920's and did not achieve wide circulation until after the foundation of the journal *Electronics* by McGraw-Hill Publishing Co. in 1930. (The older meaning also survives: as an example, the Institute of Physics and Physical Society in Britain has an Electronics Group concerned with physical electronics.)

Electronics in the technological sense has been variously subdivided into categories that encompass most of the profession of electrical engineering. For instance, when the American Institute of Electrical Engineers and the Institute of Radio Engineers combined in 1962 to form the Institute of Electrical and Electronics Engineers, the new

organization comprised 34 specialized groups, most of them pertaining to electronics, as follows:

Aero-Space	Geoscience Electronics
Aerospace and	Human Factors in
Navigational	Electronics
Electronics	Industrial Electronics
Antennas and	and Control Instru-
Propagation	mentation
Audio	Industry and General
Automatic Control	Applications
Biomedical Engineer-	Information Theory
ing	Instrumentation and
Broadcast and Tele-	Measurement
vision Receivers	Magnetics
Broadcasting	Microwave Theory
Circuit Theory	and Techniques
Communication	Military Electronics
Technology	Nuclear Science
Component Parts	Power
Computer	Product Engineering
Education	and Production
Electrical Insulation	Reliability
Electromagnetic	Sonics and Ultrasonics
Compatibility	Space Electronics and
Electron Devices	Telemetry
Engineering Manage-	Vehicular Communica-
ment	tions
Engineering Writing	
and Speech	

Such grouping points up the increasing penetration of electronics into various branches of technology and science, a process that is by no means as yet complete. Few educated people remain unaware of the role that electronics has played not only in the communications industry and in the development of computers (both with regard to military and to space applications), but also in industrial instrumentation and control (automation). Less generally appreciated is the part that electronics is coming to play in such diverse fields as food technology, geophysical exploration, medical instrumentation, materials processing, and a host of other endeavors not usually associated with electronics.

Research scientists depend on electronics *qua* technology to a surprisingly large extent. Many of the scientist's conventional instruments have been replaced by electronic devices of greater capabilities. Even more important, the advent of electronics has sparked the development of entire new branches of science such as microwave spectroscopy, electron microscopy, radio astronomy, and other fields that depend on processes such as amplifying weak signals; transmitting, recording, and analyzing large amounts of rapidly acquired data at high rates; and investigating phenomena at frequencies extending from the audio to the visible range.

Electronics is certainly repaying its large debt to physics, in numerous ways. Not only have cathode-ray oscilloscopes, vacuum tubes, and scintillation counters been joined by transistors

and other semiconductor devices to provide an ever-increasing range of complex instruments, but the very devices that have been developed as a result of certain discoveries in physics have, in turn, been used in physics research in a very direct way. A few instances will suffice to illustrate the point.

Discovery of the physical phenomenon of secondary emission of electrons has led to the development of the *photomultiplier*, an instrument that has revolutionized nuclear physics by its ability to enhance otherwise imperceptible signals and to make it possible to register extremely large numbers of discrete events.

Photoelectronic *image intensifiers*, developed in the first instance for television camera tubes, have found application in astronomy, to increase the efficiency of optical instruments and to detect faint objects against relatively bright backgrounds; in nuclear physics, to record the passage of fast nuclear particles through transparent phosphor blocks and perhaps to detect Čerenkov radiation from fast particles passing through a dense medium; in spectroscopy, to enhance ultraviolet radiation that could be otherwise detected through glass or quartz windows only with difficulty; and in the detection of x-ray patterns, both where it is necessary to detect radiation down to the shot-noise level of the x-ray quanta and in medical radiology, where the patient must be protected from excessive dosage. The applications related to nuclear physics depend on a further property of the device—its ability to be switched on or off within nanoseconds, so that one specific event out of a multitude of uninteresting events can be recorded.

The high-resolution *electron spectrometer*, derived from the electron microscope (which in turn is a descendant of the cathode-ray tube), has been used to investigate the characteristic losses of electrons transmitted through thin films of metals and alloys, providing information of basic importance for metallurgy.

Various physical principles of importance to electronics, and the corresponding devices, are described elsewhere in this book. Additional information on electronics is available in the writer's "The Encyclopedia of Electronics" (Reinhold 1962), a companion volume to the present work.

CHARLES SUSSKIND

Reference

Susskind, C., "The Encyclopedia of Electronics," New York, Reinhold Publishing Corp., 1962.

Cross-references: ANTENNAS, CAPACITANCE, CIRCUITRY, DIELECTRIC THEORY, DIODE (SEMICONDUCTOR), ELECTRICITY, ELECTROMAGNETIC THEORY, INDUCTANCE, MICROWAVE TRANSMISSION, OSCILLOSCOPES, PHOTOMULTIPLIER, RECTIFIERS, SEMICONDUCTORS, SERVOMECHANISMS, THERMIONICS, TRANSISTOR, WAVEGUIDES.

ELEMENTARY PARTICLES

The search for the elementary constituents of matter is as old as physics itself, but any quantitative attempt at such a theory had to await the experimental discoveries of this century. Such a search is prompted by two considerations: the identification of the basic building blocks of nature and the hope that their laws of interaction would be essentially simple.

The atom had to yield its claim to be indivisible when it was found that electrons were constituents of all atoms; moreover, the electrons from various species of atoms were identical. Light, with its particle properties, seemed another universal entity connected with matter, since the photons (light quanta) which were emitted in atomic transitions appeared identical apart from their momenta. The electron and the photon were the first two elementary particles to be discovered, and a quantitative theory of the emission and absorption of photons by the electrons in an atom was possible only after the invention of quantum mechanics. The corresponding picture of the atom regarded the electrons in an atom as being subject to electrostatic attraction of the positively charged nucleus (and the mutual repulsion of the electrons), the photons having only a transitory existence being either emitted or absorbed in the transitions between the atomic states. The search for the structure of matter now became a search for the constituents of the nucleus.

A quantum theory of the nucleus (or rather nuclei) was made possible by the discovery of the proton and the neutron. The nuclear interaction which was responsible for holding the nucleus together (against the disruptive electrostatic repulsion of the protons) was found to be of an entirely new kind, much stronger than the electric interaction at short distances but decreasing very much more rapidly with distance. The various complex nuclei differ in the number of protons and neutrons they contain.

By that time, the theory of the interaction between electrons and photons had developed to the point where the electrostatic repulsion or attraction between electrically charged particles could be understood in terms of the exchange of photons between them. In the lowest nontrivial approximation, it gave the Coulomb law for small velocities. The basic interaction was the emission and absorption of "virtual" photons by charged particles. A similar mechanism could be invoked to explain the short-range nuclear interaction; and essentially our present picture of the nuclear interaction is that it is due to the exchange of particles, which have nonzero masses which are a fraction of nuclear mass—the same approximation procedure used for deducing the static Coulomb force (from the electron-photon interaction) is no longer valid here; and the nuclear force has a rather complicated form. However, these theoretical considerations did predict the existence of a set of three particles called pions, which have since been discovered.

Another kind of particle and another kind of

interaction were discovered from a detailed study of beta radioactivity in which electrons with a continuous spectrum of energies are emitted by an unstable nucleus. The interactions could be viewed as being due to the virtual transmutation of a neutron into a proton, an electron and a new neutral particle of vanishing mass called the neutrino. The theory provided such a successful systematization of beta decay data for several nuclei that the existence of the neutrino was "well-established" more than twenty years before its experimental discovery. The beta decay interaction was very weak even compared to the electron-photon interaction.

Meanwhile, the electron was found to have a positively charged counterpart called the positron; the electron and positron could annihilate each other, with the emission of light quanta. The theory of the electron did in fact "predict" the existence of such a particle. It has, since then, been found that the existence of such "opposite" particles (antiparticles) is a much more general phenomenon (see below).

Our present catalog of elementary particles and decay modes contains many more entries. These particles fall into five families: the photon family, the electron family, the muon family, the meson family and the baryon family. Most of these particles are unstable and decay within a time which is often very small by normal standards but which is many orders of magnitude larger than the time required for any of these particles to traverse a typical nuclear dimension. There is a wide variety of reactions between them, but they could be understood in terms of three basic interactions—the strong (or nuclear) interaction⁵, the electromagnetic interactions and the weak interactions. The nuclear forces and the interaction between pions and nucleons belong to the first, the electron-electron and electron-photon interactions to the second, and the beta decay interaction to the third. The present theoretical framework enables us to handle more or less quantitatively the electromagnetic and weak interactions only. Despite this, it is possible to understand many aspects of strong (as well as the other) interactions in terms of conservation laws and invariance principles (see CONSERVATION LAWS AND SYMMETRY).

The classical conservation laws of energy, momentum and angular momentum are valid in the relativistic quantum theory of elementary particles also. The particles may possess intrinsic angular momentum or "spin" which expressed in natural units $\frac{h}{2\pi}$ of angular momentum is restricted to an integer or a half-odd integer. Angular momentum conservation holds only when this spin angular momentum is included. But one finds that to every particle, there corresponds an antiparticle with the same mass, same spin and same lifetime. (In the case of the photon and the neutral pion, they are their own antiparticles; they are strictly neutral particles.) The particle and antiparticle have equal and opposite electric

charges, and the antiparticle of the antiparticle is the original particle. The conservation of electric charge is another familiar (although non-classical) conservation law satisfied by all known interactions. It is the prototype of a set of "additive" conservational laws which include the conservation of the baryon number, the muon number and the electron number. To the best of our knowledge, these conservation laws are still exact.

In addition to these additive conservation laws which arise from continuous symmetries, there are a set of "multiplicative" conservation laws which are associated with discrete symmetries. It is possible to examine the invariance of the physical laws under space inversion, i.e., using a left-handed coordinate system instead of a right-handed coordinate system or vice versa; if the statement of the law is unaffected by this interchange, it is possible to show that a quantum number having the two values ± 1 can be assigned to classify the quantum-mechanical states such that a state with the label $+1$ will not change to one with the label -1 due to any interaction. This quantum number is called "parity." Just as particles may possess intrinsic angular momentum (spin), particles may also have intrinsic parity. Table I lists the particles (and their corresponding antiparticles) with their respective additive quantum numbers, intrinsic parities and lifetimes. General principles of relativistic quantum theory imply that antiparticles of integral spin particles have the same parity as the particles; for half-odd-integral spin particles the antiparticle has the parity opposite that of the particle. All experimental checks are in accordance with this prediction.

In addition to invariance under space inversion, we may consider particle conjugation (replacement of particles by antiparticles), invariance, and time reversal invariance, or combinations of these transformations. It turns out that strong and electromagnetic interactions are invariant under each of these three transformations (and hence any product of these), but weak interactions are invariant only under combined inversion (product of particle conjugation and space inversion) and under time reversal. It can be shown that all interactions are invariant under the product of the three transformations of space inversion, particle conjugation and time reversal if some very general principles of the relativistic quantum theory of these particles are valid.

One notices that the various particles belonging to a family have the same spin and the same values of the additive quantum numbers except the electric charge. The photon has a universal interaction with all charged particles; it has been found possible to connect the conservation of electric charge and this universal interaction structure, on the one hand, to the vanishing mass and unit spin of the photon, on the other. The electron and muon partake of both electromagnetic and weak interactions, but do not exhibit any strong interaction. In fact the muon family appears to be simply a duplicate of the electron family except for a change in the unit of mass. At

the present time, no basic reason has been found for this doubling, and it is perhaps the most fascinating puzzle of current elementary particle physics. The members of these two families are collectively known as leptons and they all have spin one-half. The neutral members (the electron neutrino, the muon neutrino, and their antiparticles) have extremely weak interaction with matter since they do not participate even in electromagnetic interactions.

The meson family consists of seven members which fall into a triplet of pions, a doublet of kaons and a doublet of antikaons. They are all pseudoscalar (spin zero and odd parity) and exhibit strong interactions. The charged particles are of course coupled to the photon, but even the neutral members can participate in electromagnetic interaction by virtue of the large probability for virtual dissociation into charged particles. They participate in a variety of weak interactions including the nuclear beta decay interaction.

It is found that the kaons, the hyperons (baryons other than the neutron and proton) and their antiparticles, collectively known as "strange particles," can decay by weak interactions not involving leptons or photons with a lifetime which is large compared to the natural periods appropriate to strong interactions. On the other hand, these particles are produced copiously in high-energy nuclear collisions. These two circumstances can be understood in terms of the existence of another additive quantum number, called hypercharge, which is conserved in strong and electromagnetic interactions but violated in weak interactions.

The meson-baryon system exhibits further regularities as far as strong interactions are concerned. The neutron and the proton have very nearly the same mass and similar nuclear interactions although their electromagnetic properties are quite different. The three pions have different electric charges, but again they have approximately equal masses and similar nuclear interactions. This kind of multiplet structure is evident for other strongly interacting particles: the kaons form a doublet, the sigma hyperons form a triplet, and the xi hyperons form a doublet, and the lambda hyperon remains a singlet. In view of the relative weakness of the electromagnetic interaction, it is tempting to ascribe all deviations from exact equality of the masses to the indirect action of the electromagnetic interaction. In this framework, it is possible to consider the members of a multiplet to be different states of the same particle corresponding to the values of a new quantum number. What is remarkable is that if one takes this point of view, it is possible to show that the strong interactions exhibit a remarkable invariance under a group of continuous transformations which may be viewed as the group of rotations in a fictitious three-dimensional space (or more correctly as the special unitary group SU_2 of transformations on two variables). The transformations act as follows: the singlet is unchanged, the doublet components

TABLE I. CATALOG OF ELEMENTARY PARTICLES

Particle	Family*	Spin	Mass (MeV)	Lifetime† (sec)	Antiparticle	Parity‡	Charge	Hypercharge	Baryon Number	Electron Number	Muon Number
Photon, γ	Photon	1	0	Stable	Photon, γ	-	0	0	0	0	0
Electron neutrino, ν_e	Electron	$\frac{1}{2}$	0	Stable	Antielectron neutrino, $\bar{\nu}_e$	Undefined	0	0	0	1	0
Electron, e^-		$\frac{1}{2}$	0.51098	Stable	Positron, e^+	+	-1	0	0	1	0
Muon neutrino, ν_μ	Muon	$\frac{1}{2}$	0	Stable	Antimuon neutrino $\bar{\nu}_\mu$	Undefined	0	0	0	0	1
Muon, μ^-		$\frac{1}{2}$	105.66	2.21×10^{-6}	Positive muon, μ^+	+	-1	0	0	0	0
Neutral pion, π^0	Meson	0	135.0	3×10^{-16}	Neutral pion, π^0	-	0	0	0	0	0
Positive pion, π^+		0	139.6	2.55×10^{-8}	Negative pion, π^-	-	+1	0	0	0	0
Neutral kaon, K^0		0	497.8	1.0×10^{-10}	Neutral antikaon, \bar{K}^0	-	0	+1	0	0	0
Positive kaon, K^+	Baryon	0	493.9	6×10^{-8}	Negative antikaon, \bar{K}^-	-	+1	+1	0	0	0
Proton, p		$\frac{1}{2}$	938.2	Stable	Antiproton, \bar{p}	+	+1	+1	+1	0	0
Neutron, n		$\frac{1}{2}$	939.5	1.01×10^{-3}	Antineutron, \bar{n}	+	0	+1	+1	0	0
Lambda, Λ		$\frac{1}{2}$	1115.4	2.2×10^{-10}	Antilambda, $\bar{\Lambda}$	+	0	0	1	0	0
Positive sigma, Σ^+		$\frac{1}{2}$	1189.4	$.8 \times 10^{-10}$	Negative antisigma, $\bar{\Sigma}^-$	+	+1	0	1	0	0
Neutral sigma, Σ^0	Baryon	$\frac{1}{2}$	1193.2	1×10^{-11}	Neutral antisigma, $\bar{\Sigma}^0$	+	0	0	1	0	0
Negative sigma, Σ^-		$\frac{1}{2}$	1197.6	1.6×10^{-10}	Positive antisigma, $\bar{\Sigma}^+$	-	-1	0	1	0	0
Neutral Xi, Ξ^0		$\frac{1}{2}$	1316	1.5×10^{-10}	Neutral antiXi, $\bar{\Xi}^0$	-	0	-1	1	0	0
Negative Xi, Ξ^-		$\frac{1}{2}$	1321	1.3×10^{-10}	Positive antiXi, $\bar{\Xi}^-$	+	-1	-1	1	0	0

* Electron and muon families are collectively known as the lepton family. The proton and neutron are both nucleons; other members of the baryon family are the hyperons.
† The neutral kaon has a long-lived component K_S^0 and a short-lived component, which are quantum mechanical superpositions of the neutral kaon and the neutral antikaon.
‡ Electron, muon, proton, neutron and lambda parities are defined by convention.

transform like the components of a spinor, and the triplet components transform like the components of a vector. This property of strong interactions is called "charge independence"; and the corresponding conserved dynamical variable (with three components) is called the isotopic spin. It then turns out that hypercharge conservation is a consequence of isotopic spin conservation and electric charge conservation. While the conservation of isotopic spin is violated by the electromagnetic (and weak) interactions, the charge independence of nuclear interactions is still expected to be satisfied to within a few per cent and experimental tests confirm this. Since the symmetry associated with invariance under isospin transformations is not directly related to space-time properties, one often refers to it as an "internal symmetry."

One might now raise the question: Which of these particles are basic constituents of matter? For the case of the atom, say the simplest of them all, the hydrogen atom, it seems easy to say that it is a composite system made up of an electron and proton bound together by an electrostatic force. However, this answer is not completely satisfactory since the electrostatic force itself is due to the exchange of light quanta, and in the process of atomic transitions photons are emitted or absorbed. Yet we do not include them as constituents of the atom. In beta radioactivity, electrons and neutrinos emerge from the nucleus, yet the nucleus is not pictured as containing either of these varieties of particles but rather as made up of protons and neutrons. The beta electron and neutrino are rather assumed to be created at the moment of emission. With the mesons taking part in strong interactions, however, such distinctions are no longer obvious, and the question of whether a particle is elementary or is composed of several other particles cannot be answered except perhaps within the context of a more quantitative but limited model. A point of view that has gained some acceptance is that *none* of these particles are elementary and that each is a composite of several particles (including perhaps itself)!

This view, while by no means inevitable or even well-established, is a possible picture, because in the realm of elementary particles we can not only add particles together to construct a composite system, we can also "subtract" particles by adding antiparticles. The claim that particles A and B go to make up the particle C is difficult to distinguish from the claim that particles \bar{B} (antiparticle to B) and C go to make up the particle A. Further, particles play a dual role. On the one hand, they are constituents of a composite system; on the other hand, they are the objects which are exchanged to generate forces between the constituents. In any case, in view of the very large number of entries in Table 1, it is not desirable to accept all of them as the ultimate constituents of matter.

This is even more forcefully brought to our attention by the recent discovery of a very large number of ultra short-lived particles. They appear as sharp resonances in multiparticle systems.

Since these "resonances" disintegrate within a short time (even on the nuclear scale!), it is difficult to view them as elementary particles, but they seem to play an important role in interaction phenomena and are produced as often as the more stable (and familiar) mesons and baryons included in Table 1. It appears at the present time that they ought to be included on more or less the same footing. A list of the better established resonances is given in Table 2 along with the mesons and baryons from Table 1. Since an unstable particle lives only for a very short time, its energy and consequently its mass cannot be sharp, and from elementary quantum mechanical considerations we would expect this "width" in the mass of a resonance to be inversely proportional to its lifetime. Since the width is what is measured experimentally, the width (rather than the lifetime) is usually quoted in connection with resonances.

Since these particles are coupled in the strong interactions, one would expect them to occur in isospin multiplets. This is in fact observed. It turns out that since strong interactions are invariant under particle conjugation and are charge independent, we could define a multiplicative quantum number called G-parity which has definite values ± 1 for mesons and meson resonances. These values are also included in Table 2.

One notes also that the meson and baryon multiplets seem to fall into further supermultiplets. Following the analogy of the isospin group, we may now ask what internal symmetry group is responsible for this interaction. We must also remember that whatever is responsible for the violation of this higher symmetry must itself be a part of the strong interaction. A scheme in which invariance under the special unitary group on three variables (SU_3) holds approximately has been successful in correlating and predicting the spectrum of particles and their interactions. The isospin group, SU_2 , is a subgroup of this unitary group. Just as for charge independence, no basic reason has been found for the origin of this "unitary symmetry." Still other symmetry groups, even wider than SU_3 and generally incorporating it, and which are even more significantly violated, are being studied. It appears that the complete understanding of these higher internal symmetries would involve not only their origin, but also the origin of their violation.

The theoretical description of the interaction of electrons and photons involves a relativistic quantum theory of the electron and photon fields (QUANTUM ELECTRODYNAMICS). While fundamental difficulties still remain, the computational techniques developed have enabled the accurate prediction of even the finer details of the electron's electromagnetic properties. Such a method of calculation completely fails for strong interactions, although there are reasons to believe that the correct theory should involve interacting quantized relativistic fields. The immediate correspondence between each kind of particle and a field as obtained in quantum electrodynamics is unlikely for the strongly interacting particles. On the other hand the answer to the question "What

TABLE 2. STRONGLY INTERACTING PARTICLES AND RESONANCES (AS OF 1964 SUMMER)

Particle or Resonance	Spin	Mass (MeV)*	Width (MeV)+	Parity	Electric Charge	Hyper- charge	Isotopic Spin	G-parity	Unitary Symmetry Assignment
Pion, π	0	138.5	0	-	0, +1	0	1	-	Pseudoscalar meson octet
Kaon, K	0	495.8	0	-	0, -1	+1	$\frac{1}{2}$	Undefined	
Antikaon, \bar{K}	0	495.8	0	-	0, 1	-1	$\frac{1}{2}$	Undefined	
Eta, η	0	548	10	-	0	0	0	+	Baryon octet
Nucleon, N	$\frac{1}{2}$	938.9	0	-	0, -1	+1	$\frac{1}{2}$	Undefined	
Lambda, Λ	$\frac{1}{2}$	1115.4	0	-	0	0	0	Undefined	
Sigma, Σ	$\frac{1}{2}$	1193.4	0	-	0, +1, 1	0	1	Undefined	Vector meson octet
Xi, Ξ	$\frac{1}{2}$	1318.4	0	-	0, 1	1	$\frac{1}{2}$	Undefined	
Rho resonance, ρ	1	750	106	+	0, +1, -1	0	1	+	
Kaon resonance, K^*	1	888	50	-	0, -1	-1	$\frac{1}{2}$	Undefined	Vector meson singlet
Antikaon resonance, \bar{K}^*	1	888	50	-	0, 1	-1	$\frac{1}{2}$	Undefined	
Phi resonance, ϕ	1	1020	3	-	0	0	0	-	
Omega resonance, ω	1	782	9	-	0	0	0	-	Baryon resonance decuplet
Nucleon resonance, N^*	$\frac{3}{2}$	1238	125	-	0, +2, -1, -1	-1	$\frac{1}{2}$	Undefined	
Y resonance, Y^*	$\frac{3}{2}$	1385	53	+	0, +1, 1	0	1	Undefined	
Xi resonance, Ξ^*	$\frac{3}{2}$	1530	8	+	0, 1	1	1	Undefined	
Omega minus resonance, Ω	$\frac{3}{2}$	1685	0	+	-1	-2	0	Undefined	

* The average mass of the members of the isotopic multiplet is tabulated.
+ Since the unstable particles of Table 1 live "practically forever" on the nuclear time scale, the corresponding widths are several orders of magnitude smaller than one MeV; these are quoted here as "0".

are the primitive fields?" is not obvious. Thus we are faced with a frustrating situation: there are reasons to believe that the basic theory should involve relativistic interacting quantized fields, but there is no immediate way of deciding the number or nature of the fields or of the law of their interaction. It is to be hoped that this situation will not persist indefinitely!

Since the quantum field theory framework has not so far yielded acceptable computational techniques for strong interactions, in recent years increasing effort has gone into attempts to make a theory of the reaction amplitude ("scattering matrix elements") directly. The original hope was that the general principles of quantum theory might give sufficient information to determine the reaction amplitudes more or less completely. While this program has not so far been successful, it has given rise to a variety of computational techniques for strong interactions. These, together with the discovery of the resonances, have enabled us to understand the qualitative (and in some instances, quantitative) features of strong interactions.

To sum up, we find that we have now a very large number of "elementary" particles which, by their very number, forfeit their claim to be considered ultimate constituents of matter. We have some understanding of the regularities observed in their spectrum and their interactions, and we have discovered a variety of conservation laws. However, we still do not understand the multiplicity of these particles, nor do we have a quantitative theory of their interactions. Perhaps yet another level of discovery awaits us in our search for the constitution of matter.

E. C. G. SUDARSHAN

References

- Fermi, E., "Elementary Particles," New Haven, Yale University Press, 1951.
 Ford, K. W., "The World of Elementary Particles," New York, Blaisdell Publishing Co., 1964.
 Marshak, R. E., and Sudarshan, E. C. G., "Introduction to Elementary Particle Physics," New York, Interscience Publishers, 1962.
 Kallen, G., "Elementary Particle Physics," Reading, Mass. Addison-Wesley Publishing Co., 1964.
 Nishijima, K., "Elementary Particles," New York, W. A. Benjamin, 1963.
 Hamilton, J., "Elementary Particles," London, Oxford University Press, 1959.
 Streater, R. F., and Wightman, A. S., "TCP Theorem and All That," New York, W. A. Benjamin, 1964.
 Chew, G. F., "S-Matrix Theory of Strong Interactions," New York, W. A. Benjamin, 1961.

Cross-references: ANTIPARTICLES; CONSERVATION LAWS AND SYMMETRY; QUANTUM ELECTRODYNAMICS; STRONG INTERACTIONS; WEAK INTERACTIONS.

ELEMENTS, CHEMICAL

The idea of simplicity underlying the bewildering complexity of nature has always been a conceptual thread underlying man's view of the

world. The Greek philosophers of antiquity were among the first to record their speculations, and we are to this day influenced in an unconscious way by their thoughts on the elements, which they supposed to be the ultimate components of matter and chemical change. Thus we speak of "man's battle with the elements" and "the raging elements" in unconscious reflection of the ideas of Thales, Anaximenes, Heraclitus and Empedocles of the fifth century B.C. who believed that all matter was made of one or more of the elements earth, air, fire and water. These ideas did not prove particularly fruitful in advancing our understanding of the nature of matter and chemical change. Nevertheless, it was not until van Helmont (1648) that they were challenged on a rational basis.

In 1662, Robert Boyle, in the "Sceptical Chymist" gave a reasonably clear definition of a chemical element with operational overtones which we can accept today. "I mean by elements, . . . certain Primitive and Simple, or perfectly unmingled bodies; which not being made of any other bodies, or mingled bodies, are the Ingredients of which all those called perfectly mixt bodies are immediately compounded, and into which they are ultimately resolved." He gave no list of elements, however, this being left to Lavoisier who published a naturally incomplete, but remarkably accurate list in 1789, in his justly famous "Traité de Chimie."

The Definition of "Element." In modern language, Boyle can be paraphrased in the following way: an *element* is a chemical species that cannot, by ordinary chemical manipulation be decomposed into a number of simpler chemical species. It is the entity that survives intact the infinite variety of transformations that a sample of matter can be caused to undergo. Every *compound* is composed of two or more of these species and can be decomposed into them by suitable chemical procedures. This definition provides an operational means of identifying an element in terms of laboratory procedure, and by it, any species that defies decompositional efforts must be classified as an element. Such an assignment must of course be somewhat tentative, since in a number of cases, substances that have stubbornly resisted decomposition and therefore carried the classification, have ultimately yielded as new techniques developed.

With the recent growth in detailed knowledge of atomic structure, it became possible to define an element in terms of the submicroscopic structure of matter. Such a definition is relieved of the ambiguities mentioned above. Thus an element is a sample of matter that consists of only one kind of atom, the atoms being identified in terms of their atomic number, or nuclear charge. Each element is composed of atoms having characteristic nuclear charge and an equivalent number of electrons. This nuclear charge can be determined by the charged-particle scattering technique first employed by Rutherford or by the simpler and more precise method of Moseley which relates the frequency of the characteristic x-rays produced

by the element upon electron bombardment to the nuclear charge z by means of the equation

$$w = R(z - b)^2$$

where w is the wave number of the x-ray, z is the nuclear charge, and b and R are constants.

This definition removes the ambiguities created by such observations as the decomposition of elemental molecular hydrogen or nitrogen by high temperatures into atomic species, or the decomposition of the rare gases into charged species in an electric discharge. This type of decomposition, which does not effect the underlying nuclear structure is thus excluded from the operational definition originating with Boyle.

Numbers and Kinds of Elements. The number of substances recognized as chemical elements has steadily increased since the publication of Lavoisier's list which included about thirty of the true elements. Today it is recognized that there are some 90 naturally occurring elements, the exact number depending upon the level of abundance considered limiting. There are also about 13 artificial ones with this latter number possibly increasing as the techniques of nuclear science improve. The artificial elements include those with atomic numbers above 92 as well as promethium and technetium. Technetium is also observed in stars, and in that sense is naturally occurring.

The 103 presently known elements represent distinct chemical species differing by integral units of positive charge (corresponding to the charge of the proton) beginning with element number 1, hydrogen, and progressing through element number 103. Species having nuclear charges between 1 and 103 have all been identified, so it can be said that the list of elements is complete, except for the possibility of adding new ones with atomic numbers greater than 103. The list of elements, however, is not expected to get much longer since the instability of the nucleus rapidly increases with high atomic number, all of the elements beyond bismuth (83) being naturally radioactive.

Most of the elements exhibit variations in mass, due to the varying numbers of neutrons present in their nuclei. Atoms having the same nuclear charge but differing in mass number or atomic weight are referred to as ISOTOPES. If all of these are considered, then there are approximately 1000 different atomic species represented in the list of chemical elements. Only 18 of the elements existing in nature exhibit a single mass number, and some, tin being a good example, have as many as ten naturally occurring stable species differing only in their neutron number or mass. All of the elements exhibit a variety of mass modifications which are artificially produced, but these are unstable or radioactive. These artificially produced species, of course, make up the bulk of the previously mentioned 1000 different entities. The *naturally occurring* radioactive isotopes, in addition to those beyond bismuth, include carbon 14, chlorine 36, vanadium 50, potassium 40, rubidium 87, indium 115, lanthanum 138, neodymium 144, samarium 147, lutetium 176, tantalum 180, rhenium 187, and platinum 190.

The Natural Distribution of Elements. The relative abundance of the elements is quite different for the earth's crust from what it is thought to be for the universe as a whole. These terrestrial abundances are listed in Table 1 from which it can be seen that only thirteen elements comprise over 98 per cent of the earth's crust, the oceans, and the atmosphere.

TABLE 1. THE THIRTEEN MOST ABUNDANT ELEMENTS IN THE EARTH'S CRUST

Element	Abundance (%)
Oxygen	49.52
Silicon	25.75
Aluminum	7.51
Iron	4.70
Calcium	3.39
Sodium	2.64
Potassium	2.40
Magnesium	1.94
Chlorine	1.88
Hydrogen	0.88
Titanium	0.58
Phosphorus	0.120
Carbon	0.087

It will be noted that many of the common and important elements of commerce are not included in this list, but rather belong to the remaining 2 per cent of the earth's crust. Copper, lead, and nitrogen are especially conspicuous by their absence.

If we turn our attention now to cosmic abundances, the list has quite a different make up (see Table 2). This list of course is known with considerably less accuracy since it is our attempt to guess at the relative abundance of the chemical elements in the entire universe: our galaxy, all the other galaxies, and the vast, but not entirely empty spaces in between. This information has largely been gathered by spectroscopic studies of the light emitted by the luminous bodies in these galaxies and by careful analysis of the samples of off-planet material (meteorites) that constantly shower the earth.

TABLE 2. THE THIRTEEN MOST ABUNDANT ELEMENTS IN THE UNIVERSE

Element	Relative Abundance*
Hydrogen	3.5×10^8
Helium	3.5×10^7
Oxygen	2.2×10^8
Nitrogen	1.6×10^8
Carbon	8×10^4
Neon	2.4×10^4
Iron	1.8×10^4
Silicon	1×10^4
Magnesium	9×10^3
Sulfur	3.5×10^3
Nickel	1.3×10^3
Aluminum	8.8×10^2
Calcium	6.7×10^2

* These abundances are relative to silicon taken as 1×10^4 .

It can readily be seen that this list is quite different from the terrestrial abundance list. For the universe as a whole, the elements hydrogen and helium far outrank all others, while on the earth, hydrogen is only tenth in abundance and helium doesn't even appear on the list. Oxygen remains high, and nitrogen, absent from the terrestrial list, is the fourth most abundant element when cosmic abundances are considered. In Fig. 1, some of the interesting variations in abundance are displayed. The relative abundance is plotted against mass number.

It will be noted that elements with atomic weights that are multiples of four and two are more abundant than nearby elements with relatively similar atomic weights. Examination shows that those elements having *proton or neutron* numbers 8, 50, 82, and 126 also exhibit maxima. These numbers are so-called magic numbers conferring an especially high degree of nuclear stability.

The Origins of the Elements. Considerable speculation has been devoted to the question

of the origin of the particular atomic weight distribution of elements thought to represent cosmic abundances. Examination of Fig. 1 suggests that this distribution is related to nuclear stability so that it would seem fruitful to consider what kind of an environment and sequence of events might lead to the observed distribution.

It should first of all be clear that in none of these theories is an attempt being made to truly consider *origins*, but only to explain the present distribution of atomic weights relative to some rather arbitrarily chosen point of time in the past. Nothing is postulated concerning the events preceding the zero time chosen.

One of the earliest and most popular theories, due mainly to Gamow, assumes that at time zero, the universe consisted of a dense mass of neutrons and radiation which at that time began a rapid expansion. During the early period, the temperature and density rapidly fell to values prevalent today. In the first five minutes, the universe was converted to a mixture of protons, neutrons, and

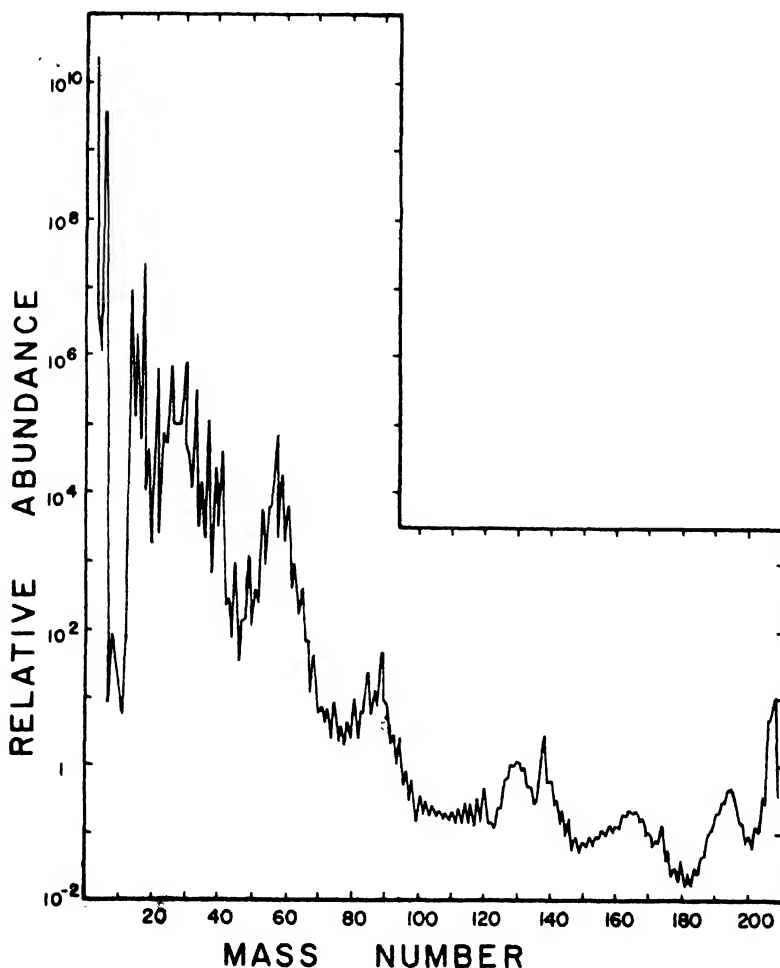


FIG. 1. [From Cameron, A. G. W., *Sky and Telescope*, 254 (May, 1963)]

electrons; in the next five minutes, sufficient cooling had occurred to permit thermonuclear reactions to occur, leading predominantly to the light elements. By the end of thirty minutes, the temperature of the expanding universe had dropped below that required for thermonuclear reactions in light elements and the distribution that we observe today, in which about 99 per cent of the universe is composed of hydrogen and helium, was achieved. The build-up of the heavier elements is somewhat less clear. This theory assumes that the buildup was generally by neutron capture, but the absence of any stable nucleus of mass 5 throws some doubt on this general mechanism.

A theory based on equilibrium arguments involving the relation of binding energies of the nuclei to their abundance, predicts a much higher initial density of matter than the preceding kinetic theory and suggests that all of the elements were formed in exploding novae. While this theory is adequate in predicting abundances up to about mass 60, it seems quite inadequate for the heavier elements.

The most recent theory of formation of the elements, which is also the most detailed, assumes a steady-state process, without beginning or end. Matter is continuously formed and condenses into new stars and galaxies which undergo a series of evolutionary changes as a result of internal nuclear reactions and eventually again distribute their matter into interstellar space.

The observational and experimental basis for this theory is mainly the *variations* in elemental distribution that are known for various types of stars in our galaxy. These variations suggest that the heavier elements, especially, are the natural consequence of stellar evolution. Detailed reaction schemes have been developed which account quite well for the high abundance of hydrogen and helium, and the low abundance of carbon, nitrogen and oxygen. The peak in mass number around 56 is also quite adequately explained. These reaction schemes suggest an age for our galaxy between 10 and 15 billion years.

RUSSELL H. JOHNSON

References

- Cherdyn'tsev, V. V., "Abundance of Chemical Elements," Chicago, University of Chicago Press, 1961 (translated by W. Nichiporuk).
 HAISSINSKY, M., "Nuclear Chemistry and its Applications," Reading, Addison-Wesley Publishing Company, Inc., 1964 (translated by D. G. Tuck).
 Johnson, R. H., and Grunwald, E., "Atoms, Molecules and Chemical Change," Second edition, Engelwood Cliffs, N.J., Prentice-Hall Inc., 1965.
 Gamow, G., "Matter, Earth, and Sky," Englewood Cliffs, N.J., Prentice-Hall, Inc., 1958.
 Cameron, A. G. W., "Birth of the Elements," *Sky and Telescope*, 254 (May 1963).
 Fowler, W. A., "The Origin of the Elements," *Chem. Eng. News*, 90 (March 16, 1964).

Cross-references: COSMOLOGY, ELECTRON, ISOTOPES, NEUTRON, PERIODIC LAW AND PERIODIC TABLE, PROTON, RADIOACTIVITY.

ENERGY. See WORK, POWER AND ENERGY.

ENERGY LEVELS

The term "energy level" is used in referring to discrete amounts of energy which atoms and molecules can have with respect to their electron or nuclear structure. The concept of permissible discrete energy levels was first introduced by Planck in explaining the physical basis for the spectral distribution of blackbody radiation. A second related principle due to Planck was that the emission and absorption of radiation are associated with transitions between these energy levels, the energy thereby lost or gained being equal to the energy, $h\nu$, of the quantum of radiation. Here h is Planck's constant and ν is the frequency of the radiation.

The first application of energy levels in the electron structure of atoms to explain optical spectra was made by Bohr. The original Bohr atom had as its basis that the only allowable states of an atom were those in which the electronic angular momentum was an integral multiple of $h/2\pi$. Circular orbits suggested by Bohr were extended by Sommerfeld to include the quantization of momentum in elliptic orbits, and to provide an improved explanation of optical spectra.

These early concepts were modified by the development of the theory of wave mechanics, in which it was shown that the allowable "stationary" states for the electrons in an atom must represent solutions of the Schrodinger wave equation. These solutions are conveniently represented by a set of "quantum numbers" for each electron. On this basis the electron structure of an atom containing any number of electrons can be built up. Two further concepts which are essential to this picture, however, are electron spin proposed by Uhlenbeck and Goudsmit, and the exclusion principle due to Pauli. In addition to the angular momentum of the electron in its orbit, each electron possesses angular momentum due to spin about an axis. The Pauli exclusion principle specifies that no two electrons in an atom can exist in the same quantum state, corresponding to the same set of quantum numbers.

Each electron in an atom can be characterized by four quantum numbers, n , l , m_l , m_s . The energy of an electron depends principally upon the positive integer n , and larger values of n correspond to larger electronic orbits. The quantum number l possesses physical significance in terms of angular momentum in the orbit, and is constrained to have the values of zero or positive integers less than n . The number, m_l , represents the component of l along a given axis, and must take on the values of zero or positive and negative integers whose absolute values are less than or equal to the value of l . The quantum number,

m_s , can be $+\frac{1}{2}$ or $-\frac{1}{2}$, and represents the component of the spin along the axis.

Within a given atom, electrons having the same value of the principal quantum number, n , form a definite group or "shell." Those electrons possessing the same value of l for a given value of n are in the same subgroup or "subshell." The possible number of electrons in a shell or subshell depends upon the possible values of m_l and m_s . Whenever a subshell is filled, the total angular momentum of the electrons involved is zero. Electrons outside of filled subshells contribute additional angular momentum which is summed vectorially and assigned the numbers J , L and S . Here J is representative of the total angular momentum, L the orbital angular momentum, and S the spin angular momentum.

An atom is stable only when it exists in the state for which the quantum numbers of its electrons give the lowest total energy. The energy of the atom may be increased to a higher level by having an electron "excited" to another state

represented by a different set of allowed quantum numbers. Transitions back again to the "ground" state will be accompanied by the emission of radiation. Wave mechanics indicates, however, that only certain transitions from one quantum state to another can be probable. These "selection rules" specify that $\Delta L = \pm 1$ and that $\Delta J = 0$ or ± 1 .

Following the principles just given, an electron energy-level diagram can be constructed for the excited states of the atoms of any particular isotope. Discrete energy levels will exist for the allowed quantum states of excited electrons. The spectrum of radiation which can be emitted from the isotope will be determined by the energy differences between these states, where the transitions involved are allowed by selection rules. Figure 1 illustrates such an energy-level diagram for sodium.

This diagram indicates that very little difference in energy level results from changes in the electron spin orientation. These orientations cor-

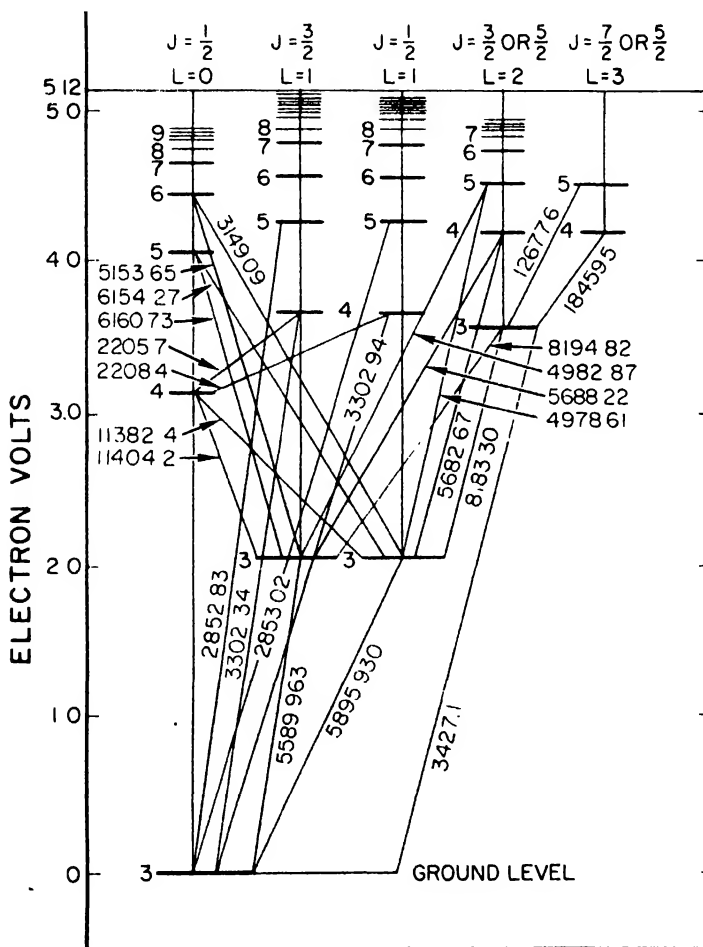


FIG. 1. Energy-level diagram for sodium. The large numbers are wave lengths in \AA of radiation emitted or absorbed during the indicated transitions. Principal quantum numbers are shown as integers.

respond to different values of J for the same value of L . No attempt is made to show a separation in the diagram for the cases of $L = 2$ and $L = 3$. Such closely spaced energy levels due to the effect of the coupling of the electron spin give rise to "fine structure" in spectra.

Many spectral lines, when examined with high resolving power instruments, are found to exhibit a still finer structure of several lines very close together. This is termed "hyperfine structure" and has been found to be due to two causes. One is the isotope effect in which atoms of different isotopes of the same element possess slightly different excited electron energy levels. The other cause of hyperfine structure has been determined to be due to the fact that the atomic nucleus also possesses angular momentum, which is vectorially added to the electronic angular momentum and quantized. Differences in the resultant states of the atom correspond to very small differences in the energy levels and hence in the observed spectrum.

Another source of structure in spectra results when the atoms emitting or absorbing the radiation are in a magnetic or an electric field. In a field, space quantization in the direction of the field takes place. The values of the magnetic moment or electric moment of the atom associated with the various possible components of angular momentum as quantized in the field direction result in different energy levels. This splitting of levels is referred to as the Zeeman effect in the case of an applied magnetic field, and the Stark effect in the case of an applied electric field (see ZEEMAN and STARK EFFECT). The amount of splitting increases with the intensity of the superposed field. In addition, the spectral structure can vary due to a tendency for the orbital momentum, L , and the spin momentum, S , to become uncoupled and undergo space quantization independently in high fields. In weak fields, the space quantization is determined from the total angular momentum J .

Any electron of an atom may be excited to some higher allowed energy level by absorption of the amount of energy specified by the difference in the energy levels involved. By absorption of a sufficient amount of energy, any electron can be removed from an atom, resulting in ionization. It is not necessary to consider only the outermost loosely bound electrons. When electrons from inner shells are excited or removed, the process of returning to the "ground" or lowest energy state involves the emission of "characteristic" x-rays. They are "characteristic" in that the x-ray spectrum produced is typical of the particular atom producing the radiation. Atoms of higher atomic number and transitions involving electrons in innermost shells produce higher-energy radiation. Absorption, as well as emission, of characteristic x-ray radiation is observed between allowed electronic energy levels. For further information on atomic spectra and energy levels, see reference 1.

In addition to the energy levels associated with particular types of atoms, wave mechanics

shows that discrete quantum states and energy levels are associated with molecular structure. New energy levels arise from vibration and rotation of molecules. To illustrate the allowable vibrational states for a diatomic molecule, consider Fig. 2, which shows the mutual force, F , and the potential energy, V , plotted as a function of the separation, r , between the two atoms. Possible energy levels are indicated by the dotted lines. Even the lowest state corresponds to an energy greater than V_0 , where the force between the atoms would be zero, and hence has some associated kinetic energy.

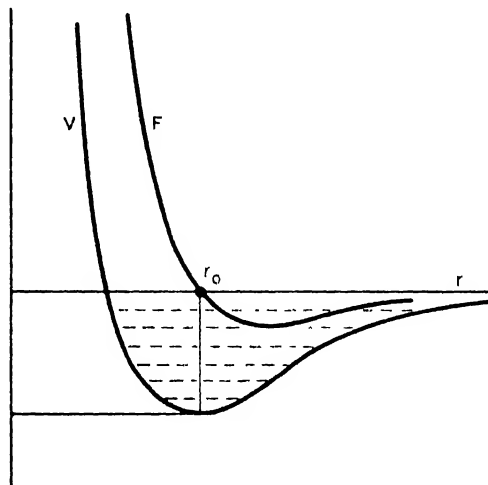


FIG. 2. Mutual potential energy V and force F as a function of atomic separation in a diatomic molecule. F is repulsive when positive and attractive when negative.

The rotation of a molecule has a quantized angular momentum which can be vectorially combined with that of the electrons. Transitions from one level to another usually involve a change in the electronic state as well as in the rotational state. Since energy differences due to allowed changes in rotational motion are very small compared to the energy differences in electronic states or vibrational states, the effect of transitions in rotational states is to produce bands of very closely spaced frequencies in the emission and absorption spectra. Such "rotational" bands are observed in molecular spectroscopy, depending upon what other transitions may be simultaneously involved, in the ultraviolet region, the visible, the infrared, and even the microwave region. See reference 2 for further details on molecular energy levels and associated spectra.

An atom may not only combine with others to form a molecule, but may be one of a large number of atoms forming a crystal. Here solutions to the wave equation show that within the solid, the individual energy levels of the free atom broaden into bands of overlapping levels. This is illustrated in Fig. 3. The bulk electrical and optical properties of solids are determined

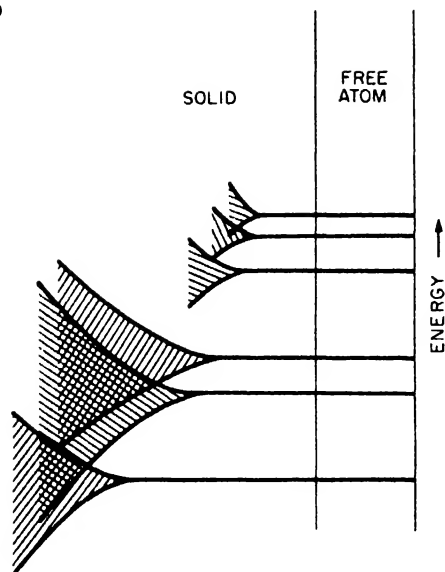


FIG. 3. The electron levels of a free atom split into bands when the atom enters a solid.

by the nature of these energy-level bands. Metals have the highest occupied energy band unfilled, while satisfying the exclusion principle that no two electrons occupy the same state. This permits electrons to gain energy under the application of an electric field and flow freely through the solid. Insulating crystals have the highest occupied energy band filled and appreciably separated from the next higher band. Semiconductors represent an intermediate situation where electrons can be injected into an unfilled band to contribute to conduction.

If the solid crystal is not completely regular but contains imperfections or impurity atoms, specific electron energy levels will be associated with these sites. This condition is responsible for luminescence and phosphorescence in certain solids. Additional information on energy levels and energy bands in solids is given in reference 3.

Finally, it should be mentioned in addition to energy levels associated with the electronic structure of atoms and the motion of atoms in molecules, there are energy levels associated with the structure of nuclei. Such levels evidence themselves in various ways, particularly through the different energies of the electromagnetic radiation, or gamma rays, emitted from excited nuclei.

WILLIAM E. PARKINS

References

1. White, Harvey E., "Introduction to Atomic Spectra," First edition, New York, McGraw-Hill Book Co., 1934.
2. Herzberg, Gerhard, "Molecular Spectra and Molecular Structure," Second edition, New York, D. Van Nostrand & Co., 1950.
3. Seitz, Frederick, "The Modern Theory of Solids," First edition, New York and London, McGraw-Hill Book Company, 1940.

Cross-references: ATOMIC PHYSICS, ELECTRON SPIN, MOLECULES AND MOLECULAR STRUCTURE, NUCLEAR STRUCTURE, SCHRÖDINGER EQUATION, SEMICONDUCTORS, SPECTROSCOPY, X-RAYS, ZEEMAN AND STARK EFFECTS.

ENTROPY

The word "entropy" was coined from the Greek by Rudolph Clausius in 1850 to mean "transformation". He applied it to a thermodynamic function, S , defined by the differential equation

$$dS = dQ/T \quad (\text{reversible}) \quad (1)$$

where dQ is the element of heat added reversibly to a system at absolute temperature T . In the classical macroscopic approach to thermodynamics, no meaning is assignable to entropy. The classical view is that thermodynamics is a machine into which some facts are put and other facts emerge.

In 1948, Claude Shannon³ demonstrated that the change in the function defined by

$$S_1 = k \sum p_i \log p_i \quad (2)$$

measures the amount of information in any message (p_i = probability that the receiver assigns to the receipt of the i th possible). The sum is over all possible messages. It has been shown that the entropy of Clausius is derivable from that of Shannon.¹ Before considering this derivation, a simple example of the meaning of entropy will be considered. Suppose you are asked to guess which number from 1 to 100 has been secretly written on a piece of paper. Taking each p_i equal to $1/100$, it is found that in the above equation, $S_1 = k \log (100)$. Letting k equal 1 and taking the logarithm to the base 2 gives $S = 6.67$ "bits." This means if each question has only two possible answers (i.e., "yes" or "no"), it will take between 6 and 7 questions to find the number. The questioner begins by asking, "Is the number between 1 and 50?" Depending on the answer, he continues by reducing the possible numbers by a factor of 2 with each question. Since $2^6 = 64$ and $2^7 = 128$, the questioner will surely find the number in 7 questions and 1/3 of the time he will find it in 6 questions. Entropy may therefore be said to measure the state of ignorance of a person relative to a well-defined question, if the person only knows a probability distribution. It measures the expected number of questions he will have to ask in order to go from his state of partial knowledge to a state in which he knows everything about the well-defined question. According to the Shannon derivation,³ S_1 represents the amount of information in a message telling the actual number written on the paper (for someone for whom $p_i = 1/100$). This interpretation of entropy is used in information theory in the design of codes and in the analysis of information transmission systems.

To make the connection with classical thermodynamics, consider a physical system and ask the question "In what quantum state is this system?" Of course one can never say in which quantum state a system resides but rather can only give a

probability for the system being in a particular state. The probability distribution for the states must be consistent with the observer's knowledge. The information theory principle of maximum entropy says to choose a set of probabilities which agrees with the available data and maximizes the entropy, for, in accordance with the meaning of entropy, this is the most noncommittal view. A state of equilibrium is, by the definition given by Gibbs,⁴ a state of maximum entropy. From the information theory point of view, it is a state in which all of the random motions which can take place are, in fact, taking place so that the observer knows as little about the systems as it is possible to know beyond his knowledge of the "constants of the motion." If we describe a system by giving only the pressure, temperature and volume (or other gross properties), we omit many details. Entropy measures how much more there is to be said before the quantum state is specified.

As an illustration, consider a closed system of particles. From quantum mechanics, we know the system is in some state, i , with energy ϵ_i . If we make an observation of ϵ , the best we can do is infer that it represents the expectation energy, $\langle \epsilon \rangle$, for this is a repeatable quantity associated with the motion. To generate the appropriate probability distribution, we maximize S_i defined in Eq. (2), with the following constraints on the p_i :

$$\sum p_i = 1 \text{ (the system is in some state)} \quad (3)$$

$$\sum p_i \epsilon_i = \langle \epsilon \rangle \text{ (the system has an expectation energy)} \quad (4)$$

Maximization of the entropy subject to the two equations given leads to the probability distribution

$$p_i = e^{-\psi - \beta \epsilon_i} \quad (5)$$

The resultant entropy, $S_{i,\max}$, is, by Gibbs' definition, the equilibrium entropy. The probability distribution is known as the Boltzmann distribution. It is easy to demonstrate that the parameter β is equal to $1/kT$ where k is the Boltzmann constant and T is the absolute temperature. If the observer is limited to macroscopic observations concerned with the energy of a body, the above derivation leads to laws which connect the observations to one another. That is, the quantities ψ , β , $\langle \epsilon \rangle$, S and various combinations are related to one another by equations from which p_i and ϵ_i have been eliminated. These relations are known as the "Laws of Classical Thermodynamics".

For example, since, from Eq. (4)

$$d\langle \epsilon \rangle = \sum_i \epsilon_i dp_i + \sum_i p_i d\epsilon_i$$

it is clear that the changes in the energy of a body may be divided into two classes: (a) those of the type $\sum_i \epsilon_i dp_i$ which necessarily change the probabilities and therefore the entropy (non-isentropic) and (b) those in which $\sum_i \epsilon_i dp_i = 0$ (i.e., isentropic). This division into two classes of energy exchange gives rise to the concepts of "heat" and "work." For a detailed account of the derivation see reference 2.

Entropy, as a logical device for generating probability distributions, has been applied in reliability engineering, decision theory, and the theory of steady-state irreversible processes. (See list of references with reference 5.) The generalized approach to entropy usage was initiated by Jaynes.⁶

M. TRIBUS

References

1. Tribus, M., "Information Theory as the Basis for Thermostatistics and Thermodynamics," *J. Appl. Mech.*, B (March 1961).
2. Tribus, M., "Thermostatistics and Thermodynamics," D. Van Nostrand Co., 1961.
3. Shannon, Claude, "A Mathematical Theory of Communication," *Bell System Tech. J.* (July and October 1948).
4. Gibbs, J. W., "Collected Works," Vol. I, p. 56, New Haven, Conn., Yale University Press.
5. Tribus, M., and Evans, R., "The Probability Foundations of Thermodynamics," *Appl. Mech. Rev.*, 765 (October 1963).
6. Jaynes, E. T., "Information Theory and Statistical Mechanics," *Phys. Rev.*, 106, 620; 108, 171 (1957).

Cross-references: BOLTZMANN'S DISTRIBUTION LAW, PHYSICAL CHEMISTRY, STATISTICAL MECHANICS, THERMODYNAMICS.

EQUILIBRIUM

In the elementary sense of the macroscopic (visible to the naked eye) system, equilibrium is obtained if the system does not tend to undergo any further change of its own accord. Any further change must be produced by external means.

Mechanical and Electromagnetic Systems. Equilibrium in mechanical and/or electromagnetic systems is reached when the vectorial summation of generalized forces applied to the system is equal to zero. In any potential field, that is, gravitational or electric vector potential, force can be expressed as gradient of potential (magnetic force however, is a curl of a vector potential). The potential energy therefore has an extremum at the equilibrium configuration. For example, a system such as a mass suspended by a string against the gravitational force (or its weight) is at mechanical equilibrium if the tensile force in the string is equal to the weight of the mass it supports. The d'Alembert principle further states that the condition for equilibrium of a system is that the virtual work of the applied forces vanishes.

Thermodynamic Systems. When a hot body and a cold body are brought into physical contact, they tend to achieve the same warmth after a long time. These two bodies are then said to be at thermal equilibrium with each other. The zeroth law of thermodynamics (R. H. Fowler) states that two bodies individually at equilibrium with a third are at equilibrium with each other. This led to the comparison of the states of thermal equilibrium of two bodies in terms of a third body called

a thermometer. The temperature scale is a measure of state of thermal equilibrium, and two systems at thermal equilibrium must have the same temperature (see THERMODYNAMICS).

Generalization of equilibrium consideration by the second law of thermodynamics specifies that the state of thermodynamic equilibrium of a system is characterized by the attainment of the maximum of its ENTROPY. Thermodynamic coordinates are defined in terms of equilibrium states.

Equilibrium between two phases of a system is reached when there is no net transfer of mass or energy between the phases. Phase equilibrium is determined by the equality of the Gibbs functions (also called free enthalpy, free energy, or chemical potential) of the phases in addition to equality of their temperatures and stresses (such as pressure and/or field intensities—intensive properties). Equilibrium of first-order phase change requires continuity of slope or first derivative of the Gibbs function with respect to an intensive property and is generalized as the Clapeyron relation. Second- and higher-order phase changes are given by the condition of continuity of curvature or second derivative of the Gibbs function and so on.

Chemical or nuclear equilibrium of a reactive system is reached when there is no net transfer of mass and/or energy between the components of a system. At chemical or nuclear equilibrium, the Gibbs function of the reactants and the products must be equal according to stoichiometric proportions, in addition to uniformity in temperature and stresses. Chemical equilibrium is summarized in the form of the Law of Mass Action. The trend for the displacement from an equilibrium state is specified by LeChâtelier's principle.

Thermodynamic equilibrium is reached when the condition of mechanical, electromagnetic, thermal, phase, and chemical and nuclear equilibrium is reached.

Stability of Equilibrium. A process or change of state carried out on a system such that it is always near a state of equilibrium is called a quasi-stationary equilibrium. This requires that the process be carried out slowly. If a mechanical system is initially at the equilibrium position with zero initial velocity, then the system will continue at equilibrium indefinitely. An equilibrium position is said to be stable if a small disturbance of the system from equilibrium results only in small, bounded motion about the rest position. The equilibrium is unstable if an infinitesimal displacement produces unbounded motion. In the gravitational field, a marble at rest in the bottom of a bowl is in stable equilibrium, but an egg standing on its end is in unstable equilibrium. When motion can occur about an equilibrium position without disturbing the equilibrium, the system is in neutral (or labile, or indifferent) equilibrium, an example being a marble resting on a perfectly flat plane normal to the direction of gravity. It is readily seen that stable equilibrium is the case when the extremum of potential is a minimum.

When dealing with general thermodynamic systems, the fact that entropy tends to a maximum in the trend toward equilibrium of a natural process generalizes the above mechanical consideration with respect to stability. An equilibrium state can be characterized as a stable equilibrium when the entropy is a maximum; neutral equilibrium when displacement from one equilibrium state to another does not involve changing entropy; and unstable equilibrium when entropy is a minimum. Any slight disturbance from an unstable equilibrium state of a system will lead to transition to another state of equilibrium.

Statistical Equilibrium. In the microscopic sense, that is, treating systems in terms of elemental particles such as molecules, atoms, and other material or quasi-particles (such as photons in radiation, phonons in solids and liquids, rotors in liquids), equilibrium states are recognized as the most probable states. An equilibrium state of a system is therefore defined in terms of most probable distributions of its elements among microscopic states which may be defined in terms of energy states. In this sense, statistical equilibrium is a condition for macroscopic equilibrium and an equilibrium state of a system is one of its extremal states. In the methods of STATISTICAL MECHANICS, the probability of distribution is expressed in terms of the density of distributions in the phase space. Based on the Liouville theorem, if a system is in statistical equilibrium, the number of the elements in a given state must be constant in time; which is to say that the density of distribution at a given location in phase space does not change with time. For an isolated system, the distribution is represented by a microcanonical ensemble. At equilibrium, no phase point can cross over a surface of constant energy, and the density of distribution is preserved. In this case individual molecules of a system can be represented by phase points. Any part of an isolated system in statistical equilibrium can be represented by a canonical ensemble. A subsystem of a large system in thermal equilibrium also behaves like the average system of a canonical ensemble. A system and a constant temperature bath together can be considered as an isolated system. A phase point in a canonical ensemble can represent a large number of molecules, thus accounting for strong interactions. A canonical ensemble is characterized by its temperature and is therefore pertinent to the concept of thermal equilibrium. When applied to equilibrium of systems involving mass exchange, such as a chemical system, we have a "particle bath" in addition to a constant temperature bath. The pertinent representation for equilibrium including mass exchange as well as energy exchange is known as a grand canonical ensemble, which accounts for the chemical potentials of its elements.

When applied to a system with a large number of elements, the distributions are measured by thermodynamic probability (W); the most probable distribution is such that W is a maximum. This optimal principle is consistent with the condition of maximum entropy (S) given in

the above. The Boltzmann hypothesis states that $S = k \ln W$, where k is the Boltzmann constant.

Depending on the specifications of W , namely, those of Maxwell-Boltzmann (for low concentration of distinguishable particles, weak interaction and high temperature, such as a dilute perfect gas), Fermi-Dirac (for elemental particles with antisymmetric wave functions at high concentrations of indistinguishable particles and low temperatures, such as electrons in metal), or Einstein-Bose (for elemental particles with symmetric wave functions, such as He^4 at high concentration of indistinguishable particles and low temperature), equilibrium distributions take different forms (see BOSE-EINSTEIN STATISTICS and FERMI-DIRAC STATISTICS). The Maxwellian speed distribution in a dilute perfect gas is a distribution based on Maxwell-Boltzmann statistics.

As a consequence of molecular considerations, when two systems are connected for transfer of mass without significant transfer of energy, such as two containers at different temperatures connected by a capillary tube, we have the relation of thermal transpiration.

Trend toward Equilibrium. The mechanism by which equilibrium is attained can only be visualized in terms of microscopic theories. In the kinetic sense, equilibrium is reached in a gas when collisions among molecules redistribute the velocities (or kinetic energies) of each molecule until a Maxwellian distribution is reached for the whole bulk. In the case of the trend toward equilibrium for two solid bodies brought into physical contact, we visualize the transfer of energy by means of free electrons and phonons (lattice vibrations).

The Boltzmann H -theorem generalizes the condition that with a state of a system represented by its distribution function f , a quantity H , defined as the statistical average of $\ln f$, approaches a minimum when equilibrium is reached. This conforms with the Boltzmann hypothesis of distribution in the above in that $S = -kH$ accounts for equilibrium as a consequence of collisions which change the distribution toward that of equilibrium conditions.

Consideration of perturbation from an equilibrium state leads to methods for dealing with rate processes and methods of irreversible thermodynamics in general.

Fluctuation from Equilibrium. A necessary consequence of the random nature of elemental particles in a body is that the property of such a body is not at every instant equal to its average value but fluctuates about this average. A precise meaning of equilibrium can only be attained from consideration of the nature of such fluctuations. In the above, we have repeatedly considered a "large" number of particles. It is important to know how large a number is "large." When considering fluctuation of energy from an average value in an isolated system, the ratio of the two is given to be proportional to $1/\sqrt{N}$, where N is the total number of elements in the system. This is also the magnitude of the fluctuation of number of particles in a system involving trans-

formation of phases and chemical and nuclear species. An equilibrium state is one at which the longtime mean magnitude of fluctuation from the average state is independent of time and this magnitude has reached a minimum value.

Large perturbation from a given state of fluctuation leads to a relaxation process toward a state of equilibrium. The relaxation time, for instance, measures the deviation from quasi-stationary equilibrium of a process which is carried out at a finite rate.

S. L. Soo

References

- Goldstein, H., "Classical Mechanics," Cambridge, Mass., Addison-Wesley Publishing Co., Inc., 1956.
Soo, S. L., "Analytical Thermodynamics," Englewood Cliffs, N. J., Prentice-Hall, Inc., 1962.

EVAPORATION. *See* VAPOR PRESSURE AND EVAPORATION.

EXCITONS

Excitons are neutral, nonmagnetic, and mobile modes of electronic excitation in insulating crystals and may be regarded as quanta of the classical polarization field in a dielectric medium. They are fundamental in the description of the interaction of light with nonmetals and play an important role in energy transport.

The original model of an exciton, developed by Frenkel in 1931, is based on the tight-binding description of a solid. In this description, applicable primarily to molecular crystals, excitons are to be identified with excited states of the constituent molecules; because of the Coulomb interaction between electrons on adjacent molecules, such excited states cannot remain localized but spread through the solid as a "wave of excitation." Frenkel's model does not take into account the possibility that an excited electron may transfer to a nearby lattice site, in effect leaving behind it a positive hole. So long as the attraction between electron and hole keeps the two from drifting apart, the resulting entity will still be a neutral form of electronic excitation, i.e., an exciton. Thus a more general picture of an exciton is that of a bound electron-hole pair; the tight-binding exciton simply has the two bound at a single site so that the separate identities of electron and hole are lost. It is not, however, strictly accurate to describe the tightly bound exciton in terms of a *single* electron-hole pair, as many electrons may participate in a given excited state. The single-pair picture works well in materials with large dielectric constant and/or small reduced electron-hole mass, particularly in semiconductors. In such solids, the electron-hole separation may attain the size of many lattice cells, and it becomes appropriate to describe the exciton as a positonium-like "atom" imbedded in a smooth dielectric medium (Wannier-Mott model). This weak-binding model is improved by

taking into account the band structure of the solid: an electron from a conduction band, characterized by parameters describing this band (symmetry, effective mass, etc.), is bound to a hole from a valence band, with the resulting exciton energy lying within the band-gap. Typically, the exciton levels are within tenths of an electron volt below the conduction band edge, out of total gap energies of several electron volts. When the electron-hole separation is of the order of a few lattice spacings, as is the case in alkali halides and solid rare gases, neither description of the exciton is quantitatively accurate, and more elaborate but less successful schemes have been devised (electron transfer model, excitation model). The great variety of properties of excitons of the various types is best illustrated by the enormous range of effective masses: these vary from as much as hundreds of electron masses for tightly bound triplet excitons (excitons in which electron and hole have parallel spin; the necessary spin flip impedes the propagation of triplet excitons) down to tenths and even hundredths of an electron mass for sums of conduction and valence band masses in some weakly bound cases.

Extensive experimental studies of excitons in molecular crystals, primarily in crystals of aromatic hydrocarbons, have been carried out as an aid in understanding the electronic structure of these materials. Of considerable importance in these studies is the existence of Davydov splitting of spectral lines: a given excited state of a molecule gives rise to as many different excitons (differing in energy, coupling to light, polarization) as there are translationally inequivalent sites in a unit lattice cell. Energy transport via excitons is investigated in these materials by observing optical spectra when foreign ("guest") molecules are imbedded dilutely in the "host" crystal; perturbation of the host material can be minimized if the impurity differs from the host only in isotopic content. It is also possible to observe collisions between excitons in pure materials; blue fluorescence has been seen emitted as the result of the annihilation of two colliding triplet excitons initially produced by red light.

The most striking evidence for the weakly bound model of an exciton is observed in hydrogen-like series of absorption lines—notably in Cu_2O , also in CdS and PbI_2 —which the simplest form of this model predicts. The model can be further probed by subjecting the excitons to electric and magnetic fields, also to strain fields. Such studies have yielded detailed understanding of much of the band structure of Cu_2O and CdS , materials in which this structure was not so accessible otherwise.

Absorption of light in insulating materials (in pure crystals, also emission) is directly related to the production (in pure crystals, also decay) of excitons. Only excitons with very small wave vectors couple to visible light since the latter has wavelengths of hundreds of lattice constants; yet the exciton wave vector is not so small as to escape detection, and the mobile character of optically produced excitons has been verified via the Lorentz

force acting on a moving exciton in a magnetic field. Another selection rule important in exciton-photon coupling comes about from the fact that the matrix element for the production of an exciton contains in lowest order a vector quantity, the transition moment μ . Only excitons with wave vector perpendicular to μ couple to light. The interaction with photons results in mixed exciton-photon modes, often called polaritons. The finite mass of the exciton part of a polariton gives rise to effects associated with spatial dispersion, e.g., anomalous modes of light propagation. Excitons, though intrinsically non-current-carrying entities, are also instrumental in photoconductivity and photoemission; the necessary charge carriers are produced from ionization of impurity states by excitons or from ionization of the excitons themselves.

To obtain bonafide absorption of light, an energy sink is needed for the excitons; except for possible re-emission at lower frequencies, phonons provide the main sink of energy. The exciton-phonon interaction may bring about the formation of "trapped excitons," which carry along a region of self-produced lattice distortion. Phonons have important effects on the line shapes of optical exciton spectra, particularly in the alkali halides where absorption peaks may attain widths of tenths of an electron volt even at low temperatures. Phonons make possible the optical production of "indirect" excitons, excitons which have wave vectors large compared to those of light. One, as yet little understood, phonon effect is the empirical Urbach law, observed in many substances, notably in the alkali halides. According to this law, absorption below the absorption edge decreases exponentially with energy, the exponent varying inversely with the temperature.

Excitons have also been studied in the x-ray region. Other related topics include vibrational excitons, propagating vibrational states in molecular crystals which are not strictly electronic in nature yet share the properties of Frenkel excitons. Excitons are also of interest in long-chain organic molecules which act like one-dimensional crystals; thus they even have biological significance.

Acknowledgments. The author wishes to acknowledge the support of the National Science Foundation and discussions with Professor J. J. Hopfield and with Dr. B. Halperin. This article was written while the author was at the University of California, Berkeley.

A. SUNA

References

Because of space limitations, we are unable to quote original references to ideas presented. The interested reader is urged to consult the references given below. Of these, reference 1 is a comprehensive review containing references to virtually all literature, up to mid-1963, relevant to semiconductors, alkali halides, and solid rare gases. For literature on molecular

excitons, see the review articles (references 2 and 3) and the exhaustive bibliographies (reference 4).

1. Knox, R. S., "Theory of Excitons," New York, Academic Press, 1963.
2. McClure, D. S., *Solid State Phys.*, **8**, 1 (1959).
3. Wolf, H. C., *Solid State Phys.*, **9**, 1 (1959).
4. Lipsett, F. R., "Energy Transfer in Polyacene Solid Solutions," I, II, III. *National Research Council of Canada Reports No. 4320, 6404; and Dielectrics*, **1**, 161 (1963).

Cross-references: PHONONS, PHOTON, POLARONS, SOLID-STATE THEORY.

EXCLUSION PRINCIPLE. See PERIODIC LAW AND PERIODIC TABLE.

EXPANSION, THERMAL

Definition and General Remarks. All substances change their shapes as a consequence of undergoing changes in temperature. A measure of this change is the thermal coefficient of expansion. For solids in the form of a thin rod or cable, this change is confined (to first order) to a change in length, and a linear coefficient of expansion is thus defined by

$$\alpha_0 = \frac{1}{L_0} \left(\frac{\partial L}{\partial t} \right)_P \quad (1)$$

where L_0 is the length of the specimen at 0°C and the subscript P implies that the pressure is kept constant. Correspondingly, for fluids and for solids of arbitrary shape, one defines a cubical or volume coefficient of expansion, β_0 , by the relation

$$\beta_0 = \frac{1}{V_0} \left(\frac{\partial V}{\partial t} \right)_P \quad (2)$$

with V_0 being the volume at the reference temperature (usually chosen to be 0°C). It may readily be shown that $\beta_0 = 3\alpha_0$. Thus for solids one usually tabulates values of α_0 while, of course, only β has meaning for fluids. Whereas β_0 is simply $1/273$ for all ideal gases (this follows from the equation $PV = nRT$), there exists a wide variation for β values among liquids. Table 1

TABLE 1. LINEAR COEFFICIENTS OF EXPANSION, α , FOR SOME SOLIDS AND CUBICAL COEFFICIENTS OF EXPANSION, β , FOR SOME LIQUIDS AT ROOM TEMPERATURE

	α (/°C)	β (/°C)
Aluminum	25.0×10^{-6}	
Copper	16.8×10^{-6}	
Nickel	12.8×10^{-6}	
Sodium	77.0×10^{-6}	
Mercury		18.2×10^{-5}
Glycerin		48.5×10^{-5}
Water (0-4°C)		-3.2×10^{-5}

gives α and β values for several substances. The negative value of β for water below 4°C is anomalous and is caused by the comparatively open

lattice structure of ice. In the case of nonisotropic crystals, the coefficient of linear expansion differs for different directions in the crystals and may even have opposite signs along different directions as is the case for CaCO_3 .

Thermodynamic Relationships. The cubical coefficient of expansion plays an important role in relating the molar specific heat at constant pressure, C_P , (which is usually measured directly in the laboratory) to the molar specific heat at constant volume (which is most often obtained from theory). This relationship, based solely on the laws of thermodynamics is¹

$$C_P - C_V = (\beta^2 VT)/X \quad (3)$$

where T is the absolute temperature and X the compressibility defined by

$$X = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T \quad (4)$$

To obtain this result we write the first law of thermodynamics (see THERMODYNAMICS):

$$\delta Q = dU + P dV \quad (5)$$

Since

$$C_V = \left(\frac{\delta Q}{\delta T} \right)_V = \left(\frac{\partial U}{\partial T} \right)_V \quad (6)$$

$$dU = \left(\frac{\partial U}{\partial V} \right)_T dV + \left(\frac{\partial U}{\partial T} \right)_V dT \quad (7)$$

and

$$dV = \left(\frac{\partial V}{\partial P} \right)_T dP + \left(\frac{\partial V}{\partial T} \right)_P dT \quad (8)$$

we obtain

$$\delta Q = \left[\left(\frac{\partial U}{\partial V} \right)_T + P \right] \left(\frac{\partial V}{\partial P} \right)_T dP + \left\{ C_V + \left[\left(\frac{\partial U}{\partial V} \right)_T + P \right] \left(\frac{\partial V}{\partial T} \right)_P \right\} dT \quad (9)$$

Thus

$$C_P = \left(\frac{\delta Q}{\delta T} \right)_P = C_V + \left[\left(\frac{\partial V}{\partial T} \right)_P \right] \left[P + \left(\frac{\partial U}{\partial V} \right)_T \right] \quad (10)$$

By the second law of thermodynamics, we have

$$dS = \frac{dU}{T} + \frac{P}{T} dV \quad (11)$$

On using Eq. (7) again this becomes

$$dS = \frac{1}{T} \left[\left(\frac{\partial U}{\partial V} \right)_T + P \right] dV + \frac{1}{T} \left(\frac{\partial U}{\partial T} \right)_V dT \quad (12)$$

Since

$$dS = \left(\frac{\partial S}{\partial V} \right)_T dV + \left(\frac{\partial S}{\partial T} \right)_V dT \quad (13)$$

one obtains on comparing Eqs. (12) and (13)

$$\left(\frac{\partial S}{\partial V}\right)_T = \frac{1}{T} \left[P + \left(\frac{\partial U}{\partial V}\right)_T \right] \quad (14)$$

$$\left(\frac{\partial S}{\partial T}\right)_V = \frac{1}{T} \left(\frac{\partial U}{\partial T}\right)_V \quad (15)$$

Since

$$\frac{\partial}{\partial T} \left(\frac{\partial S}{\partial V}\right)_T = \frac{\partial}{\partial V} \left(\frac{\partial S}{\partial T}\right)_V \quad (16)$$

we get from Eqs. (14) and (15)

$$\frac{1}{T} \frac{\partial^2 U}{\partial V \partial T} = \frac{1}{T^2} \left[P + \left(\frac{\partial U}{\partial V}\right)_T \right] + \frac{1}{T} \left[\left(\frac{\partial P}{\partial T}\right)_V + \frac{\partial^2 U}{\partial T \partial V} \right] \quad (17)$$

or

$$\left(\frac{\partial U}{\partial V}\right)_T = T \left(\frac{\partial P}{\partial T}\right)_V - P \quad (18)$$

Inserting Eq. (18) into Eq. (10), yields

$$C_P - C_V = T \left(\frac{\partial V}{\partial T}\right)_P \left(\frac{\partial P}{\partial T}\right)_V \quad (19)$$

Now

$$\left(\frac{\partial P}{\partial T}\right)_V \left(\frac{\partial T}{\partial V}\right)_P \left(\frac{\partial V}{\partial P}\right)_T = -1 \quad (20)$$

Thus on replacing $(\partial P/\partial T)_V$, the final form of Eq. (19) becomes

$$C_P - C_V = \frac{T \left(\frac{\partial V}{\partial T}\right)_P^2 \nu^2}{\left(\frac{\partial V}{\partial P}\right)_T \nu^2} \quad (21)$$

or Eq. (3) when the definition of compressibility [Eq. (4)] is employed.

Thus for a substance where β and χ are experimentally known C_V may be established from a measurement of C_P .

Grüneisen Relation.² Grüneisen introduced the parameter $\gamma = \beta V/\chi C_V$ and on the basis of simple models reached the conclusion that γ is independent of temperature. This implies that the thermal expansion coefficient is proportional to the specific heat and has the same type of temperature dependence. This is true for many substances and has, in fact, been employed as a means for predicting values for C_V at low temperatures. To illustrate the physical basis of the Grüneisen relation, we will work with a crystal model of N oscillators of identical frequency, ν , and each having equilibrium energy, ϵ . In the region of $h\nu \ll kT$, the free energy is given by:

$$F = N \left(\epsilon + 3kT \log \frac{h\nu}{kT} \right) \quad (22)$$

Since this must be a minimum at equilibrium, we obtain:

$$N \frac{d\epsilon}{dV} = -3NkT \frac{d(\log \nu)}{dV} \quad (23)$$

To obtain the thermal expansion coefficient, one expands $d\epsilon/dV$:

$$\frac{d\epsilon}{dV} = \left(\frac{d\epsilon}{dV}\right)_{V=V_0} + (V - V_0) \left(\frac{d^2\epsilon}{dV^2}\right)_{V=V_0} \quad (24)$$

But the compressibility is given by

$$\frac{1}{\chi_0} = NV_0 \left(\frac{d^2\epsilon}{dV^2}\right)_{V=V_0} \quad (25)$$

and thus

$$\frac{V - V_0}{V_0} = -3NkTX_0 \frac{d(\log \nu)}{dV} \quad (26)$$

Differentiating with respect to T and realizing that $3NkT$ is the thermal energy, we obtain:

$$\frac{\alpha V_0}{C_V \chi_0} = - \frac{d(\log \nu)}{d(\log V)} \quad (27)$$

More exact crystal models yield values for

$$\gamma_i = \frac{d(\log \nu_i)}{d(\log V)}$$

where ν_i is the i th frequency of a set of normal modes of vibration.³ These and other refinements give rise to modifications of the simple Grüneisen theory.

Source of Thermal Expansion. The dynamical basis for thermal expansion is the presence of an anharmonic component for the interaction potential.⁴ Taking x as the displacement of a lattice atom from its equilibrium neighbor separation, the potential energy has the form

$$V(x) = cx^2 - gx^3 \quad (28)$$

Then \bar{x} , the average displacement using the Boltzmann distribution function becomes

$$\bar{x} = \frac{\int_{-\infty}^{\infty} x e^{-V(x)/kT} dx}{\int_{-\infty}^{\infty} e^{-V(x)/kT} dx} \quad (29)$$

Assuming that the anharmonic term is much less than the harmonic contribution, we expand $V(x)$ to yield

$$\int_{-\infty}^{\infty} x e^{-1/kT} dx = \int_{-\infty}^{\infty} e^{-cx^2/kT} \left[x + \frac{gx^3}{kT} \right] dx \quad (30)$$

and

$$\int_{-\infty}^{\infty} e^{-V/kT} dx = \int_{-\infty}^{\infty} e^{-\frac{cx^2}{kT}} dx \quad (31)$$

Both integrals are readily evaluated to give

$$\bar{x} = \frac{3kTg}{4c^2} \quad (32)$$

or a constant temperature coefficient for thermal expansion. This simple derivation may be amplified to include specific interaction forces for the atoms of a lattice.⁵

Thermal Expansion and Curie Temperature. The Curie-Weiss law for both ferroelectrics and ferromagnets may be shown to be connected with the thermal expansion coefficient⁶. Consider a region where the dielectric constant, ϵ , is large. If we let N be the cell density and A a constant, the Clausius-Mosotti formula is

$$\frac{\epsilon - 1}{\epsilon + 2} = AN \quad (33)$$

Differentiating with respect to T , the temperature, and making use of the assumption that $\epsilon \gg 1$, one finds that

$$\frac{3d\epsilon}{\epsilon^2 dT} = \frac{1}{N} \frac{dN}{dT} = -\beta \quad (34)$$

On integrating between T and θ , the Curie temperature, one obtains

$$\epsilon = \frac{3/\beta}{T - \theta} \quad (35)$$

which is a typical Weiss law and indicates the Curie constant is of the order of the reciprocal of the thermal expansion. The special electronic structure of ferromagnetic materials also gives rise to anomalous thermal expansion coefficients in the transition region. For some materials β values are depressed and for others β values

increase more rapidly with temperature. Both magnetostriction and the variation of the energy of magnetization with the atomic size account for the anomalous behavior of different substances.

JACOB NEUBERGER

References

1. Any text in thermodynamics such as: Zemansky, M. W., "Heat and Thermodynamics," New York, McGraw-Hill Book Company, Inc., 1951.
2. Mott, N. F., and Jones, H., "The Theory of the Properties of Metals and Alloys," New York, Dover Publications, 1936.
3. Arenstein, M., Hatcher, R. D., and Neuberger, J. "Equation of State of Certain Ideal Lattices," *Phys. Rev.*, **13**, No. 5, 2087-2093 (1963).
4. Kittel, C., "Introduction to Solid State Physics," New York, John Wiley & Sons, 1956.
5. Peierls, R. L., "Quantum Theory of Solids," New York, Oxford Press, 1955. (Reprinted in 1964.)
6. Dekker, A., "Solid State Physics," Englewood Cliffs, N.J., Prentice Hall 1957; Sinnott, M., "The Solid State for Engineers," New York, John Wiley and Sons, 1958.

Cross-references: BOLZMANN'S DISTRIBUTION LAW, DIELECTRIC THEORY, HEAT CAPACITY, MAGNETISM, THERMODYNAMICS.

F

FALLOUT

The term fallout is now generally used to refer to particulate matter that is thrown into the atmosphere by a nuclear process of short time duration. Primary examples are nuclear weapon debris and effluents from a nuclear reactor excursion. The name fallout is applied both to matter that is aloft and to matter that has been deposited on the surface of the earth. Depending on the conditions of formation, this material ranges in texture from an aerosol to granules of considerable size. The aerodynamic principles governing its deposition are the same as for any other material of comparable physical nature that is thrown into the air, such as volcanic ash or particles from chimneys. However, the radioactivity associated with fallout has caused it to be put into a distinct category. This radioactivity exists because the fallout producing processes are nuclear fission and nuclear fusion.

The topographic distribution of fallout is divided into three categories called local (or close-in), tropospheric (or intermediate), and stratospheric (or world wide) fallout. No distinct boundaries exist between these categories. The distinction between local and tropospheric fallout is a function of distance from detonation to point of deposit, while the distinction between tropospheric and stratospheric fallout depends primarily on the place of injection of the debris into the atmosphere, above or below the tropopause. Whether radioactive debris from a nuclear weapon becomes tropospheric or stratospheric fallout depends on yield, height, and latitude of burst (the height of the tropopause is a function of latitude).

Because air acts as a viscous medium, a drag force is developed to oppose the gravitational force that acts on airborne particulate matter. This makes the velocity of fall dependent on particle size. The larger particles (diameters greater than about 20μ) have a higher rate of settling and create local fallout. Smaller particles injected below the tropopause are carried by prevailing winds over large regions of the surface of the earth and create the tropospheric fallout. Tropospheric fallout particles larger than about $0.1\text{-}\mu$ diameter continually mix through the circulating air mass that is in contact with the surface of the earth and gradually settle to the ground, or are washed down by rain or snow. Many smaller

particles form nuclei for raindrops. Parts of the tropospheric fallout may remain in the atmosphere for a month or more, long enough to circle the earth several times. The mean residence time above the tropopause of stratospheric fallout is from 5 to 30 months, during which time it completely encircles the earth. It gradually returns through the tropopause, primarily in certain regions where mixing between the two layers is more probable.

The exact characteristics of the radiation associated with fallout depend on the nature of the nuclear processes from which its radioactivity originates. Generally these radioactive nuclides are fission products formed from the fissioning of uranium or plutonium, but, under appropriate circumstances, considerable quantities of radioactivity can be formed through nuclear reactions induced by neutrons that are produced by the weapon or reactor. The radiation problems associated with local fallout are usually those of high-intensity gamma-ray radiation fields resulting from the relatively large quantities of radioactive material that falls back to earth within a few tens of miles from the point of origin. The important radioactive materials consist in this case of short-lived fission products and neutron-induced radioactive nuclides. The hazards of worldwide fallout come more from the problems of the long-lived radionuclides, such as Cs^{134} , Cs^{137} , and Sr^{90} , that can enter the human food chain and ultimately be absorbed by the body.

For a nuclear weapon burst in air, all materials in the fireball are vaporized. Condensation of fission products and other bomb materials is then governed by the saturation vapor pressures of the most abundant constituents. Primary debris can combine with naturally occurring aerosols, and almost all of the fallout becomes tropospheric or stratospheric. If the weapon detonation takes place within a few hundred feet of (either above or below) a land or water surface, large quantities of surface materials are drawn up or thrown into the air above the place of detonation. Condensation of radioactive nuclides in this material then leads to considerable quantities of local fallout, but some of the radioactivity still goes into tropospheric and stratospheric fallout. If the burst occurs sufficiently far underground, the surface is not broken and no fallout results.

C. SHARP COOK

References

- Israel, H., and Krebs, A., "Nuclear Radiation in Geophysics," pp. 136-240, New York, Academic Press, 1962.
- Eisenbud, M., "Environmental Radioactivity," Chs. 13 and 14, New York, McGraw-Hill Book Company, 1963.
- Cook, C. S., "Initial and Residual Ionizing Radiations from Nuclear Weapons," in Attix and Tochilin, Fds., "Radiation Dosimetry," Vol. II, New York, Academic Press (in press).

Cross-references: ATOMIC ENERGY, FISSION, FUSION, ISOTOPES, NUCLEAR REACTIONS, RADIOACTIVITY.

FARADAY EFFECT

This effect was the first demonstration of a connection between magnetism and light. Faraday found in 1845 that when plane polarized light was transmitted through glass in a direction parallel to an applied magnetic field, the plane of polarization was rotated. Since Faraday's original discovery, the phenomenon has been observed in many solids, liquids, and gases. It is important in the interaction of electromagnetic radiation with the ionosphere and in the study of charge carrier behavior in the laboratory. These effects may all be regarded as acting on the electric field associated with the wave. There is a second class of Faraday rotation phenomena in which the effect acts on the magnetic-field component of the wave. These effects are very large in ferromagnetic insulators and have made possible the development of a class of nonreciprocal microwave devices, which are described briefly in the concluding paragraph. Cgs units are used throughout this article.

Electric-field Effects. Optical Rotation from Spectral Transitions. In the region of the spectrum close to an absorption line, the dielectric properties of the medium are dominated by the absorption and its associated dispersion. In those cases of interest, the absorption line is broadened or split by an applied magnetic field. One can show by quite general arguments that for every spectral component displaced linearly in the applied field, there must be a second component that is displaced in the opposite direction. The simplest case is one in which the original spectral line splits into two lines with frequencies

$$\omega^{\pm} = \omega_0 \pm \omega_L, \quad \omega^{\pm} = \omega_0 \pm \omega_L,$$

where $\omega_L = g\hbar/4mc$ is the Larmor frequency and g is called the spectroscopic splitting factor. One can again show by general arguments that in an isotropic medium, a positively circularly polarized wave will be absorbed only at ω^+ and similarly a negative wave only at ω^- . With this theoretical background, we may now consider the mechanism of rotation. Let us imagine that a linearly polarized wave at frequency ω is directed into the medium along the magnetic field. We can expect that if we decompose the incident

wave into two waves of opposite circular polarization, the waves propagate independently. Each wave is characterized by its own dielectric constant and, therefore, its own phase velocity. If the dielectric constant of the positive component is larger than that of the negative component at frequency ω , then the positive component has the lower phase velocity and is rotated through a larger angle on passing through the medium. If we recombine the waves after a path length l , we will find the plane of polarization to be rotated through an angle

$$\theta = \pi l(n^+ - n^-)/\lambda_0$$

where λ_0 is the free-space wavelength and n^+ and n^- are the refractive indices for the two polarization directions. It is occasionally inconvenient to decompose the incident field into circular components. One can alternatively characterize the medium by a dielectric tensor. The off-diagonal elements of the tensor are given by

$$\epsilon_{xy} = \epsilon_{yx} = (\epsilon^+ - \epsilon^-)/2j$$

By comparison with the earlier expression, the rotation angle may be written approximately as

$$\theta = j\pi\epsilon_{xy}l/n\lambda_0$$

For the particular case considered here we have

$$\epsilon_{xy} = 2j\omega\omega_p^2\omega_L/(\omega_0^2 - \omega^2)^2$$

where $\omega_p = (4\pi Ne^2/m)^{1/2}$ is the plasma frequency. One of the important advantages of Faraday rotation as a technique for studying optical spectra is that it permits a determination of the Larmor frequency under circumstances where it may not be possible to resolve the splitting directly. This technique is also of very considerable importance in the determination of internal magnetic fields in ferromagnetic materials.

Carrier Rotation. One can apply an analysis similar to the above in the case of free charge carriers. It is convenient to discuss the rotation by charge carriers in two limits. At frequencies low compared with the collision frequency of the carriers and to first order in the magnetic field, we obtain

$$\epsilon_{xy} = 4\pi\sigma_0\mu H/j\omega_c$$

where $\mu = e\tau/m$ is the carrier mobility and σ_0 is the low-frequency conductivity. The low-frequency Faraday effect and the Hall effect ($q.v.$) are closely related. At high frequencies the rotation is proportional to

$$\epsilon_{xy} = j\omega_c\omega_p^2/\omega^3$$

where $\omega_c = eH/mc$ is the cyclotron frequency. The Faraday effect is particularly important in the study of semiconductivity, where carriers move with an effective mass that may be considerably different from their free mass.

Magnetic-field Effects. In addition to the high-frequency Faraday effect, which is usually associated with electric dipole transitions, there are a number of low-frequency rotation phenomena, which are associated with magnetic dipole transitions. Here the rotation is associated with the tensor properties of the magnetic permeability. The rotation angle is given by

$$\theta = -j\pi\mu_{xy}/n\lambda_0$$

Paramagnetic Rotation. In a paramagnetic material the rotation is associated with the paramagnetic resonance absorption in much the same way that optical rotation is associated with optical absorption. For a paramagnetic material, the rotation is proportional to

$$\mu_{xy} = 4j\pi\chi_0\gamma H/\omega$$

where γ is the magneto-mechanical ratio.

Ferromagnetic Rotation. In a ferromagnetic material, very large rotations may be achieved because of the very large effective susceptibility $\chi_0 \gg M/H$. For magnetically saturated material we obtain

$$\mu_{xy} = 4j\pi\gamma M/\omega$$

Antiferromagnetic Rotation. Antiferromagnetic materials are characterized by resonance absorption even in the absence of a field. Applying the magnetic field along the symmetry axis we obtain

$$\mu_{xy} = 8j\pi\chi_0\omega\gamma H\omega_0^2/(\omega_0^2 - \omega^2)^2$$

where ω_0 is the zero-field resonance frequency and χ_0 is the susceptibility perpendicular to the axis.

Applications. The principal application of the Faraday effect to electronics has been in the development of nonreciprocal microwave devices. Because of the nonsymmetric character of the permeability tensor in a magnetic field, it is possible to fabricate devices that permit the nearly unattenuated passage of microwaves in one direction but will effectively block microwave transmission in the reverse direction. These devices, which make use of ferromagnetic insulators like ferrite or garnet, are of reduced effectiveness at high frequencies because of their inverse dependence on ω . Attention is being given to antiferromagnetic materials for use at high frequencies because of their increased transverse permeability in this range.

A. M. PORTIS

Cross-references: HALL EFFECT AND RELATED PHENOMENA, LIGHT, MAGNETISM, POLARIZED LIGHT, PROPAGATION OF ELECTROMAGNETIC WAVES, SEMICONDUCTORS.

FEEDBACK

The notion of "feedback" is central to the theory of information and control, christened by the late mathematician N. Wiener in 1948 as "cybernetics," and *a fortiori* it occupies a pivotal position in the technical aspects of "automation," the latter term being coined by the American industrialist D. S. Harder in 1936. To date

"feedback," the return of "output" signals to the "input" of any device for the purpose of correcting or improving the characteristics of the device, has had greater influence within technology than within science. Within the design philosophy of modern systems, whether analogue or digital, it has become a criterion for the physical "richness" of those systems; systems not utilizing "feedback" being too simple.

The concept of the "ultrastability" of a dynamical system, i.e., the additional provision in a classical dynamical system to alter the kinds of its characteristics or to adapt to its environment, makes crucial use of secondary "feedback" to adjust other primary "feedbacks." The idea of "ultrastability," conceived by W. R. Ashby, the other principal pioneering cyberneticist, makes frequent appearance in the study of models to simulate and realize intellectual activities such as checker playing on electronic machines and theorem proving on electronic machines.

Perhaps, the oldest and simplest scheme employing "feedback" is the question-and-answer giving of ordinary conversation. If only one party does all the talking, or, more generally, all of the transmitting of signs, then there is no guarantee that the alleged message has been received, or if so, whether or not it has been understood. By continually monitoring one another, critical scientific conversation becomes possible.

Other simple, but more technical, examples of "feedback" for automatic control occur in the governor of steam engines, in thermostats, and in the automatic volume controls of radios. When more steam is fed to a steam engine its shaft accelerates. Automatic control can be realized by using a centrifugal device to tend to close off steam with greater shaft velocities. With such a scheme of "negative feedback" the engine acquires stability, i.e., it runs at an optimal velocity determined by the characteristics of the governor. On the other hand, if the centrifugal device had been installed (incorrectly) so as to supply more steam with greater shaft velocities, then this scheme of "positive feedback" would produce an instability that would either destroy the steam source or the steam engine, or both. Similarly, the automatic volume controls of radios provide "negative feedback" of a continuous or analogue variety. However, the thermostat provides "negative feedback" of a continual or discrete variety. If the local temperature exceeds a certain value, the furnace is turned off, and if the local temperature is below a certain value the furnace is turned on. Thus, unlike a steam engine and governor, the thermostat is an "on-off" device.

During and after the 1950's, even the digital computers have become components of complex sampled-data, feedback control systems. A basic idea behind sampled-data systems is that statistical sampling of signals, rather than continuous sampling, is sufficient in many cases for reliable control systems. Indeed, by the so-called Sampling Theorem, if time sampling is often enough, but not necessarily continuous, and there is an

upper bound to the electrical frequencies considered, then there is *no* loss of information.

Finally, the presently available models for the study of sequential switching circuits, which includes important parts of digital computers, all have explicit provisions for the "feedback" of informational signals.

ALBERT A. MULLIN

References

- Wiener, Norbert, "Cybernetics," New York, The M.I.T. Press and John Wiley & Sons, 1961.
 Wiener, Norbert, "The Human Use of Human Beings," Garden City, N.Y., Doubleday & Co., 1954.
 Ashby, W. R., "An Introduction to Cybernetics," New York, John Wiley & Sons, 1956.
 de Latil, P., "Thinking by Machine," Boston, Houghton Mifflin Co., 1957.
 Truxal, J. G., "Control System Synthesis," New York, McGraw-Hill Book Co., 1955.
 Fogel, L. G., "Biotechnology," Engelwood Cliffs, N.J., Prentice-Hall, 1963.

Cross-references: CIRCUITRY, COMPUTERS, CYBERNETICS, MECHANICS.

FERMI-DIRAC STATISTICS AND FERMIONS

Solid metals are good conductors of heat and electricity because about one electron per atom is free to migrate through the volume of the conductor. These electrons were once thought to behave like gas molecules which obey Maxwell-Boltzmann statistics in which the number of particles at higher energies falls off exponentially according to a relation of the form

$$n_{E,T} = \frac{1}{e^{E/kT}}$$

where E is the energy, k the Boltzmann Constant and T the absolute temperature. This electron gas theory was qualitatively useful in explaining many metallic properties, but it was never quantitatively successful. One notable failure was its prediction that electrons should contribute to the specific heats of metals.

Bohr had shown that the electron in hydrogen is not free to assume any energy, but it is restricted to certain permitted energies called quantum states or energy levels. When this quantum view of atomic electron structure was extended to more complex atoms, it was found that electrons obey the Pauli exclusion principle—only two electrons in any one atom having oppositely directed spin can occupy the same energy state. Thus in an atom with many electrons, no more than two can have the lowest permitted energy, no more than two may have the next higher permitted energy, etc. An unexcited atom with all its electrons in their lowest possible energy states includes many electrons whose energy is well above the energy of

the lowest two. The old electron gas theory of metals recognized that the inner electrons associated with each atom were quantized but assumed that the electrons that were not bound to particular atoms were entirely free to migrate through the metal with no *a priori* restrictions on their energy. Fermi-Dirac statistics describes the behavior of the electron gas under the assumption that *all* electrons within the conductor have their energies quantized and obey the Pauli principle. This new viewpoint leads to a distribution of electron energies according to a relation of the form

$$n_{E,T} = \frac{1}{e^{(E-E_f)/kT} + 1}$$

If the metal is at high temperature, this function approaches the Maxwell-Boltzmann distribution. We can see this by noting that if $T \rightarrow \infty$, the exponent of e approaches zero regardless of E . Thus, in both cases, the number of electrons of each energy tends to become uniform. The high-temperature electrons have so many states available to them that quantum restrictions make little difference. If we let the temperature approach absolute zero, the difference between these distributions becomes extreme. If T is very small, the Maxwell-Boltzmann distribution is strongly dependent on E with most particles having low E and few having high E . Indeed for $T \rightarrow 0$, the number of particles with $E > 0$ becomes zero in a Maxwell-Boltzmann gas all particles come to rest at absolute zero. A Fermi-Dirac gas behaves very differently at absolute zero. The exponent of e is plus or minus infinity depending upon whether E is greater or less than E_f . The exponential term is either infinity or zero. The denominator is either infinity or one. All energy states below E_f are filled whereas all those above E_f are empty. Thus, consistent with the assumptions, at absolute zero the electrons do not crowd into one state of zero energy but are uniformly distributed among those states which are below the critical energy E_f called the Fermi energy or the Fermi level. Fermi energies depend on the kind of metals but they are of the order of several electron volts. Thus, even at absolute zero, some electrons have energies which would be typical of a Maxwell-Boltzmann electron gas only if that gas were at several thousand degrees.

The contrast may be dramatized by the following analogy. If grains of sand are spilled on an open floor, they will spread out so they are only one deep and each has zero potential energy. If the grains are poured into a drinking straw, the straw will fill to a certain height and some grains will have considerable potential energy.

Heating a metal from absolute zero to room temperature adds only .025 eV to the average energy of its particles. Since the electrons already have a *much* greater average energy, heating a metal has but a slight effect on the energy distribution of the electrons. This accounts for the fact that electrons make a negligible contribution to the specific heats of metals, and it also explains why metals must be glowing hot before electrons

acquire enough additional energy to escape from the metal surface as in the filaments of radio tubes. Since the quantum view of electrons in a metal provides both a qualitative and quantitative picture of many metallic properties, we know metallic electrons are quantized Fermi particles rather than unquantized Maxwell particles. The application of Fermi-Dirac statistics to semiconductors accounts for their special properties as demonstrated by transistors.

From the standpoint of wave mechanics, all particles which are confined in any way are quantized. Those whose spin is integral have symmetric wave functions and do not obey the Pauli principle. If they are so numerous that they must be treated statistically, they are called *bosons* and are described by Bose-Einstein statistics. Photons are the most common bosons. Those particles whose spins are odd multiples of $\frac{1}{2}$ have antisymmetric wave functions and obey the Pauli principle. They are called *fermions* and obey Fermi-Dirac statistics. Although electrons are the most common example, protons, neutrons, and μ -mesons are all fermions with spin $\frac{1}{2}$. At high temperatures, the quantum nature of both bosons and fermions becomes insignificant and both obey the classical statistics of Maxwell-Boltzmann. The technique of deriving these distributions is called statistical mechanics.

To convey the over-all method of STATISTICAL MECHANICS, we note that it is a probability theory in which the basic technique is to compute the number of possible ways in which a system can arrange itself subject to restrictions as to the number and total energy of the particles. These ways are all assumed equally likely. (There are $52!$ shuffles of a pack of playing cards. Each is equally likely.) Then, depending on the nature of the particles, bosons or fermions, the number of distinguishable ways is computed. (In the game of bridge, there are many fewer deals $52!/(13!)^4$, than there are shuffles because the order in which a player receives his cards does not change his "hand.") The probability of any particular distinguishable distribution is proportional to the number of ways in which it can be achieved. (If we flip a coin five times, there are $2^5 = 32$ orders in which the coin can fall. Of these, there are ten ways to get two heads and only one way to get five heads. We therefore find getting two heads ten times more probable than getting five heads.) The actual expected distribution is the one which can be achieved in the largest number of ways.

JAMES A. RICHARDS, JR

Reference

Leighton, Robert B., "Principles of Modern Physics," New York, McGraw-Hill Book Co., 1959.

Cross-references: BOLTZMANN'S DISTRIBUTION LAW, BOSE-EINSTEIN STATISTICS AND BOSONS, ELECTRON SPIN, STATISTICAL MECHANICS.

FERMI SURFACE

In simple terms, the Fermi surface of a metal, semi-metal, or semiconductor is that surface in momentum space which separates the energy states which are filled with free or quasi-free electrons from those which are unfilled. Such a surface exists simply because the electrons obey Fermi-Dirac statistics. It is a surface of constant energy and is sometimes called the Fermi level.

Consider first an elementary model of a metal consisting of a lattice of fixed positive ions immersed in a sea of conduction electrons which are free to move through the lattice. Every direction of electron motion is equally probable. Since the electrons fill the available quantized energy states starting with the lowest, a three-dimensional picture in momentum coordinates will show a spherical distribution of electron momenta and, hence, will yield a spherical Fermi surface. In this model, no account has been taken of the interaction between the fixed positive ions and the electrons; indeed the only restriction on the movement or "freedom" of the electrons is the physical confines of the metal itself.

A short derivation starting with the Schrödinger equation shows that the total energy of an electron (and thus also its kinetic energy) is given by

$$E = \hbar^2 k^2 / 2m = p^2 / 2m$$

where \hbar is Planck's constant divided by 2π , k is the magnitude of the electron wave vector, m is the mass of the electron, and p is its momentum. A plot of E against k is then a parabola, as shown in Fig. 1(a). The Cartesian components of those values of k which are possible solutions to the Schrödinger equation are $k_i = 2\pi n_i / L$, where the n_i 's are integers and L is a physical dimension of the metal. Since for each energy value so defined there are actually two states (one for an electron with spin up, one with spin down), it can be shown that the density of energy states available to the electrons is

$$g(E) = \frac{(2m)^{3/2}}{2\pi^2 \hbar^3} E^{1/2}$$

where $g(E) dE$ is the number of states in the energy range E to $E + dE$. Then $n(E)$, the number of electrons per unit volume occupying energy states in this energy range, is

$$n(E) dE = g(E) f(E) dE$$

where $f(E) = \{\exp[(E - E_F)/kT] + 1\}^{-1}$, a function characteristic of particles which obey Fermi-Dirac statistics. In this expression, T is the absolute temperature, k is Boltzmann's constant, and E_F is a parameter depending on the number of electrons involved and indeed turns out to be the Fermi energy. E_F can be evaluated by integrating $n(E) dE$ from $E = 0$ to $E = \infty$ and recognizing that the integral is equal to N , the total number of electrons per unit volume. The result (at $T = 0$ K) is

$$E_F = \frac{\pi^2 \hbar^2}{2m} \left(\frac{3N}{\pi} \right)^{2/3}$$

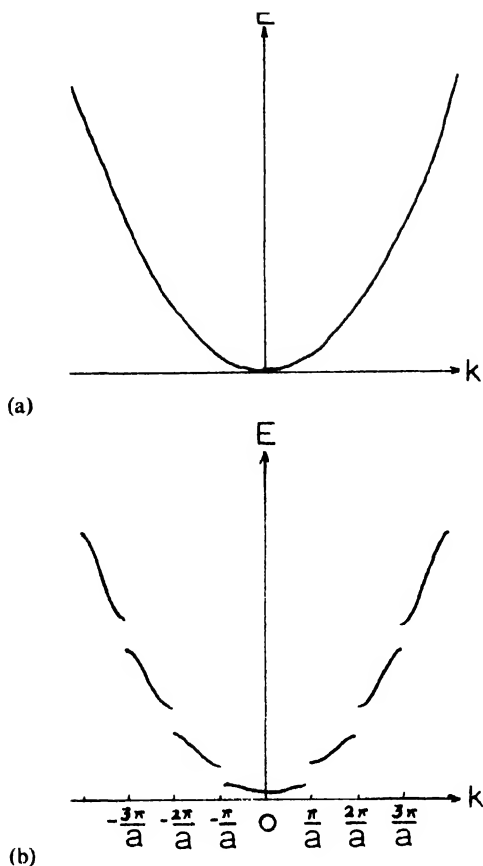


FIG. 1 (a) Energy plotted against wave number for the free electron model. (b) Energy plotted against wave number for the "quasi-free" electron model, showing energy discontinuities at Brillouin zone boundaries.

At $T = 0^\circ \text{K}$, for $E < E_F$, $f(E) = 1$, while for $E > E_F$, $f(E) = 0$. Physically this means that the probability of a state below the Fermi level being occupied is one; whereas for states with $E > E_F$ the occupancy probability drops abruptly to zero. For temperatures greater than absolute zero, the occupancy probability drops smoothly from 1 to 0 in a range of energy of width approximately equal to kT . This shell of partially filled states gives rise to the following definition: The Fermi level is the energy level at which the probability of a state being filled is just equal to one half.

A numerical evaluation of the Fermi energy for a simple metal having one or two conduction electrons per atom yields a value of approximately 10^{-11} erg, or a few electron volts. The equivalent temperature, E_F/k , is several tens of thousands of degrees Kelvin. Thus, except in extraordinary circumstances, when dealing with metals, $kT \ll E_F$; i.e., the energy range of partially filled states is small, and the Fermi surface is well defined by the statement above. It must, however, be noted that this is not necessarily true for semiconductors where the number of free electrons per unit volume may be very much smaller.

The foregoing treatment gives a qualitative insight into the physics of metals and, under some circumstances, semi-metals and semiconductors. A more detailed analysis requires that the effects of the ions in the lattice be recognized. This can be accomplished by introducing the periodic potential due to the lattice through which the electrons must move. Then the electrons are no longer "free," but, depending on the strength and character of the potentials and the approximations used in solving the Schrödinger equation, act as "quasi-free" particles. Another approach is the "tight-binding approximation"; occasionally a combination of the two approaches is used. In any case, introduction of lattice effects changes the characteristics of the model; the total energy and kinetic energy of an electron are no longer equivalent. The periodic lattice can be described conveniently in terms of Brillouin zones, each of which is large enough (in momentum space) to accommodate two electrons per atom. The Brillouin zone boundaries appear to the electrons as Bragg reflection planes or energy discontinuities, resulting in an energy versus wave number plot as shown in Fig. 1(b).

For many metals, the "nearly free" electron description corresponds quite closely to the physical situation. The Fermi surface remains nearly spherical in shape. However, it may now be intersected by several Brillouin zone boundaries which break the surface into a number of separate sheets. It becomes useful to describe the Fermi surface in terms not only of zones or sheets filled with electrons, but also of zones or sheets of holes, that is, momentum space volumes which are empty of electrons. A conceptually simple method of constructing these successive sheets, often also referred to as "first zone," "second zone," etc., was demonstrated by Harrison.¹ An example of such a construction is shown in Fig. 2. This construction works quite well, for example, for aluminum which has three valence electrons per atom. Experiments, and indeed more elegant theoretical calculations, show that the fourth zone is totally unoccupied and that the third zone monster is not multiple-connected in the manner shown. The recipe for constructing these figures, some of which may even be pleasing to art connoisseurs, cannot be developed in the limited space of this article but will be found in the references.^{1,2}

The intense research effort of the last decade on the Fermi surfaces of metals and semi-metals originated, to a great extent, with Pippard's ingenious deductions, based on anomalous skin-effect experiments, concerning the Fermi surface of copper.³ Prior to Pippard's work, it was taken for granted that in copper, with one quasi-free electron per atom, the first Brillouin zone would be only half filled and, hence, would have a nearly spherical Fermi surface. His work suggested that a series of eight necks pull out and touch the Brillouin zone boundaries in the $[111]$ crystallographic directions. This shape has now been confirmed and precisely mapped, not only for copper but also for silver and gold.

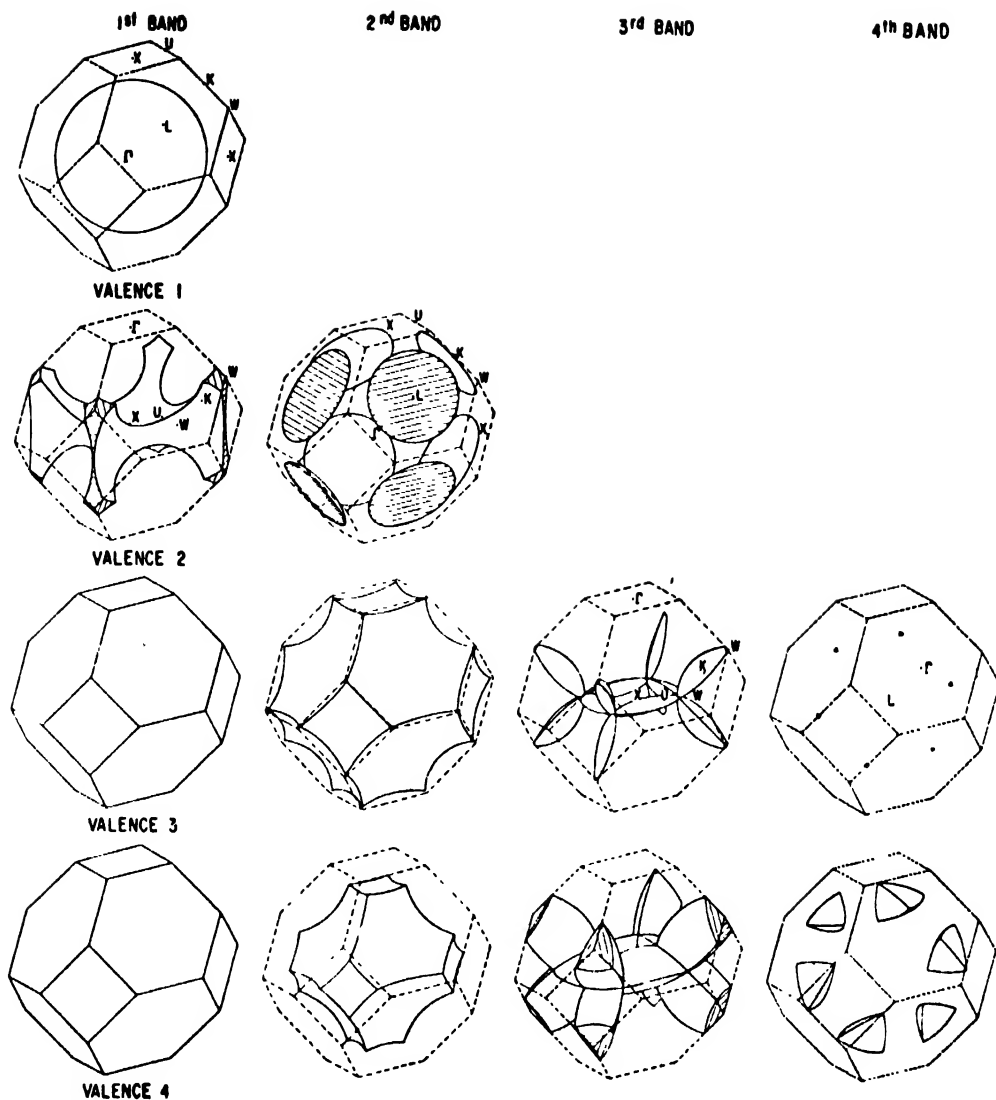


FIG. 2. Fermi surfaces in several zones or bands, for face-centered cubic metals having various numbers of "quasi-free" electrons per atom, as constructed by Harrison.¹

A variety of experimental techniques have been developed, capable of yielding both overlapping and complementary information concerning Fermi surfaces of metals and semi-metals. Listed below are some of these and the features of the Fermi surface they measure:^{1,5}

- (1) de Haas-van Alphen effect: extremal cross-sectional areas of the Fermi surface;
- (2) Cyclotron resonance: surface curvature and effective masses;
- (3) Magnetoacoustic effect: extremal linear dimensions;
- (4) Magnetoresistance: areas of contact with Brillouin zone boundaries;
- (5) Anomalous skin effect: surface curvature and surface area;

- (6) Positron annihilation: number of electrons in a given cross section;

- (7) Gantmakher effect: extremal linear dimensions.

The precision and applicability of each of these techniques is dependent on the material under investigation. Among pertinent factors are number of quasi-free electrons per atom, crystallographic structure, magnetic properties, purity, and practicality of sample preparation.

Paralleling this experimental work has been a good deal of theoretical activity. Attention has been concentrated on the calculation of Fermi surfaces for a number of metals and semi-metals, most frequently through the use of orthogonalized or augmented plane wave techniques and studies

of the theoretical bases for the interpretation of the experimental data.

H. V. BOHM
NORMAN TEPLEY

References

1. Harrison, W. A., *Phys. Rev.*, **118**, 1190 (1960).
2. Ziman, J. M., "Electrons in Metals; A Short Guide to the Fermi Surface," London, Taylor and Francis, 1963.
3. Pippard, A. B., *Phil. Trans. Roy. Soc. London Ser. A*, **250**, 323 (1957).
4. Pippard, A. B., in DeWitt, C., Dreyfus, B., and deGennes, P. G., Eds., "Low Temperature Physics," p. 3, New York, Gordon and Breach, 1962.
5. MacIntosh, A. R., *Sci. Am.*, **209**, No. 1, 110 (1963); an excellent review article for the non-physicist.

Cross-references: CYCLOTRON RESONANCE; DE HAAS-VAN ALPHEN EFFECT; ENERGY LEVELS; TRANSPORT THEORY; SOLID STATE PHYSICS; SOLID STATE THEORY.

FERRIMAGNETISM

Snoek's publication (1946) of his wartime work on ferrites established the existence of new ceramic magnetic materials capable of combining the resistivity of a good insulator (10^{12} ohm-cm) with high permeability. (see MAGNETISM.) In 1948, Néel introduced the term ferrimagnetism to describe the novel magnetic properties of these materials. A simple ferrite is composed of two interpenetrating FERROMAGNETIC sublattices with magnetizations $M_a(T)$ and $M_b(T)$ which decrease with increasing temperature and vanish at the Curie point, T_c . In a ferromagnetic material, the resulting saturation magnetization, M , would be $M_a + M_b$; however, in a ferrite, strong antiferromagnetic interaction between sublattices results in antiparallel alignment, and $M = M_a - M_b$. In general $M_a(T) \neq M_b(T)$, and the material behaves in most respects like a ferromagnet, exhibiting domains, a hysteresis loop, and saturation of the magnetization at relatively low applied magnetic fields. Practical values for saturation magnetization and Curie temperature range from 250 to 5000 oersteds and from 100 to 600 C. The high resistivity of ferrites led to early applications in low-loss coil and TRANSFORMER cores. Some ferrites display a rectangular hysteresis loop. This property has been used extensively in binary switching and memory devices.

Ferrimagnetic materials have spinel, garnet, and hexagonal structures. A typical spinel ferrite is NiFe_2O_4 . Other ferrites may be obtained by substituting magnetic (Co, Ni, Mn) or non-magnetic (Al, Zn, Cu) ions for some of the Ni or Fe ions, e.g., $\text{Ni}_{1-x}\text{Co}_x\text{Al}_2\text{Fe}_{2-x}\text{O}_4$, where x and y may be varied to modify M and T_c . Yttrium iron garnet (YIG), $\text{Y}_3\text{Fe}_5\text{O}_{12}$, is the classical ferrimagnetic garnet which combines very low magnetic loss with high resistivity. Substitution of magnetic RARE EARTH ions (Gd, Yb, Ho, etc.) for Y and of nonmagnetic ions (Ga, Al) for some

of the Fe ions leads to many different ferrite compositions with a wide range of M and magnetic loss. The rare earth ions form a third magnetic sublattice with attendant magnetization M_c antiparallel to the resultant magnetization $M_{a,b}$ of the two Fe sublattices. Since M_c and $M_{a,b}$ exhibit different variations with temperature, the net magnetization may vanish twice, at T_c and at an intermediate temperature called the compensation point, T_{comp} , where $M_c = M_{a,b}$.

A typical hexagonal ferrite is $\text{BaFe}_{12}\text{O}_{19}$. Again, other magnetic ions such as Mn, Co, and Ni may be introduced to produce wide variations in M and T_c . Hexagonal ferrites are characterized by large anisotropy fields with an axis of symmetry which may be either a direction of hard (planar ferrites) or easy (uniaxial ferrites) magnetization. They have been used as permanent magnets, in harmonic generators, and in millimeter microwave devices.

All microwave ferrite devices such as isolators, circulators, switches, phase shifters, limiters, parametric amplifiers, and harmonic generators are based on interactions of rf signals with the ferrite magnetization. Aligning M with an external biasing magnetic field, H_{dc} , and applying a microwave signal in an orthogonal direction leads to strong interaction and gyromagnetic resonance. On a microscopic scale, this is explained as application of a torque to the unpaired ELECTRON SPINS of the magnetic ions which causes them to precess at the rf frequency much like so many spinning tops. The precessional motion has a microwave RESONANCE frequency f_r dependent upon H_{dc} and the gyromagnetic splitting factor g_{eff} . In ferrimagnets with spinel structure, g_{eff} is related to the g -factors of the sublattices as follows:

$$g_{\text{eff}} = M / [(M_a/g_a) - (M_b/g_b)]$$

On a macroscopic scale, this interaction modifies the rf magnetic field in a manner which is described by introducing an antisymmetric permeability tensor $[\mu]$ whose complex components depend on M , H_{dc} , and frequency. When the frequency approaches f_r , one observes a resonance absorption line whose width, ΔH , is determined by the magnetic loss of the material. Values for ΔH cover a range from ~ 1 oersted for single-crystal YIG to ~ 1000 oersteds for some polycrystalline Ni-ferrites. The interaction of rf fields and electron spins becomes a maximum if the rf field is circularly polarized in the same sense as the precessional motion of the spins. Circular polarization in the opposite direction produces almost no interaction and no gyromagnetic resonance. This permits design of nonreciprocal ferrite devices. At high levels of microwave power, nonlinear coupling between microwave signal and precessional spin motion causes the parametric excitation of higher order modes of spin motion (magnetostatic modes and spin waves). This effect has been exploited in limiters and parametric amplifiers.

W. H. VON AULOCK

References

- Standley, K. J., "Oxide Magnetic Materials," Oxford, Clarendon Press, 1962.
 Smit, J., and Wijn, H. P. J., "Ferrites," New York, John Wiley & Sons, Inc., 1959.
 Lax, B., and Button, K. J., "Microwave Ferrites and Ferrimagnetics," New York, McGraw-Hill Book Co., Inc., 1962.

FERROELECTRICITY

Ferroelectricity is defined most reasonably as follows: When a crystal plate, in some direction, of a crystal has two stable states (different in polarization) at zero electric field and is capable of alternating between these states by means of a suitable alternating electric field, the crystal plate or the crystal is said to be ferroelectric; here, the two stable states are assumed to be identical or enantiomorphous in crystal structure and to be equal in plate thickness. In this definition of ferroelectricity, the phrase "different in polarization" is omissible, since this naturally follows from the postulate that the state transition is brought about by an electric field. If the two states were equal in polarization, the electric field would not do any work at the state transition, which means impossibility of the state transition by the electric field.

From the definition, it is concluded that any ferroelectric crystal must crystallographically belong to a noncentrosymmetric point group. Moreover, it is reasoned that the ferroelectric crystals belonging to the nonpolar (but non-centrosymmetric) groups are hardly expected to exist; in fact, none of them have yet been discovered actually. (It is not concluded that every ferroelectric crystal must belong to a polar group.)

A ferroelectric crystal is said to be regular when (1) in any crystal plate (with any Miller indices) of the crystal, the space lattice in one of the two stable states is parallel to the space lattice in the other and (2) the notion of the two stable states is possible not only for the individual crystal plates but for the crystal as a whole. From this definition, it is deduced that the regular ferroelectrics must belong to the polar groups. In general, a ferroelectric crystal belonging to a polar group satisfies the second condition for regularity if it satisfies the first condition. Therefore, the regular ferroelectrics can also be re-defined as the ferroelectrics which belong to the polar groups and satisfy the first condition, i.e., the condition of lattice parallelism. A ferroelectric crystal is said to be irregular when it is not regular. As examples, triglycine sulfate is regular and Rochelle salt is irregular.

For every ferroelectric crystal plate, it is possible to prove the existence of a spontaneous polarization P_s which must be independent of the way of choosing the reference structure (or the unit cell) and the way of emergence of the crystal-plate boundaries: P_s is equal to half the difference between the polarizations which the crystal plate shows in the two states at zero electric field.

Furthermore, for every regular ferroelectric crystal, it is possible to prove the existence of a spontaneous polarization *vector* which must be independent of the reference structure (or the unit cell), the crystal-plate orientation, and the crystal-plate boundaries. The spontaneous polarization P_s of a crystal plate, in a direction \mathbf{n} , of a regular ferroelectric crystal is equal to the component in that direction of the spontaneous polarization vector \mathbf{P}_s of the crystal: $P_s = \mathbf{P}_s \cdot \mathbf{n}$. It is only the regular ones which possess such a spontaneous polarization vector. (Existence of a spontaneous polarization is not to be supposed *a priori* in the definition of ferroelectricity, but is to be proved *a posteriori*.)

The regular and the irregular ferroelectrics are divided into fourteen and eleven kinds, respectively, in accordance with their point groups, space lattices, and types of state transition; thus, all ferroelectrics are divided into 25 kinds. (Here, the ferroelectrics belonging to the nonpolar groups are left out of consideration.) The fourteen regular kinds are denoted by the symbols

$$\begin{aligned} & \epsilon 1, rm, r2, rmm2, r4, r4mm, \\ & r6, r6mm, r3R, r3mR, \\ & r3P-I, r3P-II, r3mP-I, r3mP-II. \end{aligned}$$

The eleven irregular kinds are denoted by the symbols

$$\begin{aligned} & i1-I, i1-II, im, i2-I, i2-II, imm2, \\ & i4, i6, i3R, i3P-I, i3P-II. \end{aligned}$$

(For the meaning of each symbol see refs. 2 and 3.) If one knows the kind of a ferroelectric, one can deduce the (macroscopic and microscopic) characteristics of the ferroelectric in one state from the information about the other state and can, moreover, make some predictions on the phase transformation of the ferroelectric to a paraelectric phase.

A crystal is called paraelectric when it is not ferroelectric. This concept of paraelectricity is the broadest. Sometimes, the antiferroelectrics are separated from the paraelectrics. The notion of antiferroelectricity differs more or less from person to person. A most reasonable one is the following: A paraelectric crystal (in the general sense) is said to be antiferroelectric when by means of an electric field it can transform to another phase of whose lattice the original lattice becomes a superlattice. In this definition, one need not care about the dielectric nature of the induced phase (although usually it is regarded as ferroelectric).

KËITSIRO AIZU

References

- Aizu, K., *Rev. Mod. Phys.*, **34**, 550 (1962).
 Aizu, K., *Phys. Rev.*, **134**, A701 (1964).
 Aizu, K., *J. Phys. Soc. Japan*, **20**, 284 (1965).

Cross-references: CRYSTALS AND CRYSTALLOGRAPHY; DIELECTRIC THEORY; POLAR MOLECULES.

FERROMAGNETISM

Ferromagnetism is an example of cooperative phenomena in solids. It is characterized by a spontaneous macroscopic magnetization M (magnetic moment per unit volume) in the absence of an applied magnetic field at temperatures below a critical value known as the Curie temperature, T_c . This property is exhibited by the transition metals, Fe, Co, and Ni; the rare earth metals, Gd, Tb, Dy, Ho, Er, and Tm; and by a variety of alloys, compounds, and solid solutions involving the transition, rare earth, and actinide elements. Ferromagnetic Curie temperatures range from a fraction of a degree to hundreds of degrees Kelvin.

Cooperative magnetic behavior results from the exchange interaction between electrons which is qualitatively described as follows. Electrostatic coulomb repulsion between like electric charges acts to keep two electrons apart, a separation which is also favored by the Pauli exclusion principle if the electrons have parallel spins. Thus if two electrons are farther apart when their spins are parallel than they would be if their spins were antiparallel, the parallel state will have lower mutual electrostatic energy. However, the kinetic energies increase if the electrons are separated, and consideration of this energy may lead to lower total energy for antiparallel spins. In other words, the exchange interaction is electrostatic in nature, but is modified by details of kinetic energy and the exclusion principle, and is highly dependent on the spatial distribution of the electrons. The exchange energy between two electrons, though electrostatic, is usually expressed in the mathematically equivalent form, $-2Js_1 \cdot s_2$, where s_1 and s_2 are the spin angular momentum vectors of the two electrons. When J is positive, parallel spins represent a lower energy state than antiparallel spins. Since each electron has a magnetic moment proportional to its spin angular momentum, a state of parallel spins corresponds to a state of parallel magnetic moments.

The principal effect of the exchange interaction is embodied in the empirical Hund's rules which describe the combination of electron spins in an atom to form the atomic spin. The principal interaction between magnetic atoms, i.e., between atoms with a magnetic moment as a consequence of spin angular momentum, is thought to be of exchange character also and there has been considerable success in describing magnetic properties by assuming this interaction to have the form,

$$\mathcal{H} = - \sum_{i,j} J_{ij} \mathbf{S}_i \cdot \mathbf{S}_j.$$

Here \mathbf{S}_i and \mathbf{S}_j are the spin angular momentum vectors of atoms i and j ; the exchange integral, J_{ij} , may vary for different pairs and is usually regarded as a phenomenological parameter to be evaluated by means of experimental data. When J_{ij} is positive, parallel ordering or alignment of the atomic spins in a common direction is favored, and so there is a large spontaneous magnetization

even in the absence of an applied field. For negative J_{ij} , antiparallel spins result in lower energy (see ANTIFERROMAGNETISM, FERRIMAGNETISM).

Maximum ordering obtains at the absolute zero of temperature where the randomizing effect of thermal agitation disappears. The initial decrease of the magnetization as the temperature is increased from zero is well represented by a superposition of wavelike disturbances known as spin waves. At the Curie temperature the magnetic ordering is destroyed by thermal agitation and the spontaneous magnetization is zero (Fig. 1). Above the Curie temperature a ferromagnetic material behaves paramagnetically and has a net magnetization only in the presence of an applied field (see MAGNETISM and PARAMAGNETISM).

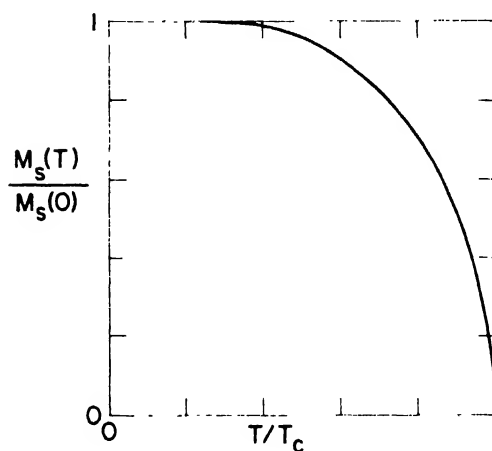


FIG. 1. Typical variation of spontaneous domain magnetization, M_s , as a function of temperature, T .

At temperatures below T_c a sample of ferromagnetic material is usually divided into small regions called domains, which vary in size and shape with a typical dimension from 0.1 to 1000 μ . Within each domain the magnetization is uniform and has the maximum or saturation value, M_s , characteristic of the temperature of the material (Fig. 1), but the direction of alignment of the individual moments in each domain changes from one domain to the next. The magnetization of the sample, which is the resultant of the magnetization of all the domains, may be much less than the saturation value (of a single domain), or it may even be zero in a completely demagnetized state.

The magnetization of a sample increases when a magnetic field, H , is applied. The value of H required to saturate the magnetization may be as small as 0.01 oersted or as large as several thousand oersteds, depending on the material. The magnetization process for a previously demagnetized material is represented by the 0-to-a portion of the magnetization curve in Fig. 2. When the applied field is subsequently

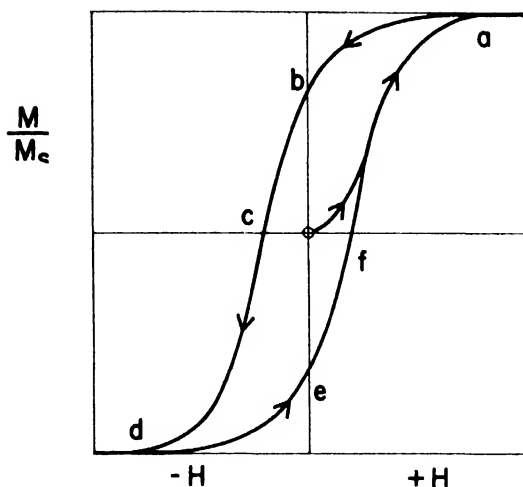


FIG. 2. Typical variation of sample magnetization, M , as a function of applied field strength, H ; the magnetization curve depends on the magnetic history of the sample.

removed, the magnetization exhibits the phenomenon of hysteresis, or lagging behind, tracing the path a-b and retaining a finite value, called the remanence, when the field is zero. The coercive force is the value of the field which must be applied opposite to the direction of the magnetic moment to trace the path b-c and reduce the magnetization to zero. Further increase of the field in this direction, followed by its removal, traces the path c-d-e. Repetition of the cycle then traces the curve e-f-a-b-c-d-e, etc. Permanent magnets are characterized by a large coercive force; i.e., a large reverse field (typically thousands of oersteds) is necessary to destroy the magnetization. However, in soft magnetic materials the coercive force may be less than one oersted.

Domains exist in a ferromagnetic material because their formation results in a lower total energy for the sample than it would have if the entire sample were a single domain. Total alignment of the magnetic moments is favored by the exchange forces, but these are usually short range forces acting between an atom and its neighbors. The dipole-dipole forces, although weaker, are long range and, alone, would orient the atomic moments like bar magnets, north pole to south pole, in closed chains to minimize the external field of the magnet (see **DIPOLE MOMENT AND MAGNETISM**). One additional factor is required for the formation of domains. This factor is anisotropy, a result of the crystal structure of most solid materials. The structure of a crystal is not the same in all directions; consequently its physical properties depend on direction. It is easier to magnetize a magnetic material in some directions, called easy axes, than in other directions. When the effect of anisotropy is superimposed on the effects of exchange and dipolar coupling, the ordered atomic moments break up

into segments or domains, so that in each domain the magnetization is uniform and lies along or near one of the easy axes. There is a large change in the direction of magnetization from one domain to the next, with the reorientation occurring gradually (on an atomic scale) in a narrow transition region known as the domain wall.

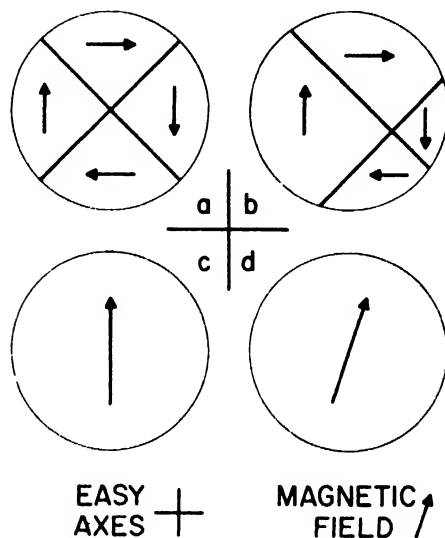


FIG. 3 Schematic representation of the change of domain structure and magnetization with applied field.

The initial part of the magnetization process, 0-a in Fig. 2, is represented schematically in Fig. 3. In the demagnetized state the domains are arranged to minimize the external field due to the magnetization and to give zero net moment for the sample (Fig. 3a). In low fields the domain boundaries move so that the domains with magnetization direction near the direction of the applied field grow in size while other domains are depleted (Fig. 3b). As the field strength is increased, the domain boundaries are swept out of the sample (Fig. 3c), and finally the (single) domain magnetization rotates into the direction of the applied field until saturation is reached (Fig. 3d). When the field is removed, domains re-form in varied orientation along the easy axes, but there is a preference for domains whose magnetization vectors lie in the easy directions nearest the direction in which the field was applied, and consequently the sample as a whole has a magnetic moment (remanent magnetization).

R. J. JOENK

References

- Bates, L. F., "Modern Magnetism," Fourth edition, Cambridge, Cambridge University Press, 1961.

Bozorth, R. M., "Ferromagnetism," New York, D. Van Nostrand Co., 1951.

Chikazumi, S., "Physics of Magnetism," New York, John Wiley & Sons, Inc., 1964.

Kittel, C., and Galt, J. K., "Ferromagnetic Domain Theory" in "Solid State Physics," Vol. 3, New York, Academic Press, 1956.

FIELD EMISSION

Field emission of electrically charged particles occurs when a sufficiently high electric field is applied to the surface of a conductor. Specifically, field emission of electrons from cold metals into a vacuum is a basic physical effect comparable to thermionic, photoelectric, or secondary emission. Field electron emission is also termed "cold emission" or "autoelectronic emission." Field emission of electrons from metals at room temperature requires an electric field of an order of magnitude of 3×10^7 volts/cm which can be obtained by applying a few thousand volts to a sharply curved cathode, which may be either a fine wire, a sharp edge, or a needle tip. From such cathodes, field emission was first observed by R. W. Wood in 1897, and technical application in high-voltage rectifiers and x-ray tubes was attempted by J. E. Lilienfeld in the early 1920's. He failed because of the inadequate vacuum techniques available at that time. The quantum mechanical theory of field emission was given by R. H. Fowler and L. W. Nordheim in 1928 which agreed with the current-voltage relationship measured by R. A. Millikan and C. C. Lauritsen, but experimental work to further verify the predictions advanced only with the possibility of controlling highly perfect emitter tips in the field emission microscope by E. W. Müller, 1937. Subsequently, field emission microscopy became an established research technique for surface phenomena connected with adsorption. By operating a point emitter at a positive potential in the presence of a gas, Müller discovered field ionization (1951) and developed the low-temperature field ion microscope (1956) which surpasses all other microscopic devices with its capability of showing the individual atoms as they constitute the crystal lattice of the metal specimen. In the last decade, technical application of field emission has also been successful with the development of powerful flash x-ray tubes (W. P. Dyke, 1955). The extremely large current densities of up to 10^8 amperes/cm² make a field emission cathode very attractive for mm wave tubes, cathode ray tubes, and electron microscopes, but the sensitivity to contamination and cathode sputtering are detrimental to stability and long lifetime.

Field emission can be explained with the concepts of quantum mechanics. The conduction electrons of a metal are moving in a potential trough, from which they can escape ordinarily only by addition of thermal energy (thermionic emission), by an energy transfer from photons (photoelectric emission), or by collision with other energetic particles (secondary emission). If the

barrier of the trough is narrowed by the application of an external electric field to be comparable with the wavelength of the electrons inside the metal, then a small amplitude of the electron wave will be noticeable outside the barrier, or in an equivalent interpretation, there is a finite probability of the electron penetrating through the potential barrier, even if its kinetic energy is insufficient to go over it. According to the Fowler-Nordheim theory of field emission, the current density J (in amperes/cm²) as a function of field strength F (volts/cm²) and of the work function ϕ (electron-volts) is approximately given by

$$1.55 \cdot 10^{-8} \frac{F^2}{\phi} e^{-\frac{6.85 \cdot 10^7 \phi}{F}}$$

A more refined theory takes into account the effect of the image force on the electron, which reduces the exponent of the above equation by a factor slightly smaller than unity and which depends upon ϕ/\sqrt{F} . The field required for a given current density is thus reduced by some 10 to 20 per cent. The temperature dependence of field emission is found to be very small below about 1000° K (Fig. 1). Considerable increase in emission is observed when both the temperature and the field are high. This effect is called T - F emission.

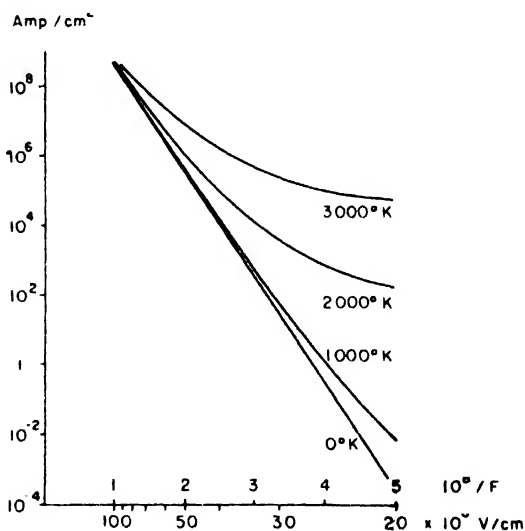


FIG. 1. Current density of field emission for a tungsten cathode plotted as a function of reciprocal field strength, for various temperatures.

Field emission of positive ions can occur when an adsorption layer is removed from the positive emitter tip by the electric field (field desorption, Müller, 1941) or at still higher field strength when the metal atoms of the emitter are coming off after single or double ionization (field evaporation, Müller, 1956). These ion currents are only transients since the source of the ions one

monoatomic adsorption layer or the tip surface itself, is soon consumed. Continuous field ion emission is obtained by operating the positive emitter in a gas of low pressure. Gas molecules attracted to the tip surface by polarization are ionized when their valence electron tunnels out into the metal while the molecule is a few angstroms above the surface. Ion currents are small (less than 10^{-8} ampere) due to the limited supply of the gas molecules, and the fields necessary for ionization are very high, about $2.5 \cdot 10^8$ volts/cm for hydrogen and $4.5 \cdot 10^8$ volts/cm for helium.

In spite of its limited technical application, field emission has been a subject of intensive investigations, and the field electron and the field ion microscopes have become productive tools of basic research in the study of metal surfaces. The specimen of the field emission microscope is a needle-shaped field emitter with a hemispherical tip of a radius of some 10^{-5} to 10^{-4} cm, arranged in a vacuum tube opposite to a fluorescent screen. With a few thousand volts applied, field emitted electrons move away from the tip in a radial direction, displaying on the screen an enlarged projection image of the distribution of electron emission at the emitter. The magnification of this microscope is approximately equal to the ratio of screen distance to tip radius and can exceed a million diameters. The lateral resolution is limited to 25 \AA by a random tangential velocity component of the electrons due to the Fermi distribution inside the metal and by diffraction due to the de Broglie wavelength. As even traces of adsorption layers change the work function and, thereby, the emission, the presence of such layers can be readily detected on the screen. The imaged tip cap usually represents a single crystal, so that the behavior of adsorption layers can be studied in its dependence on crystallographic orientation of the substrate. As little as 10^{-3} of a monolayer of oxygen is clearly discernible on many metals. Because of the small temperature dependence of field emission, such layers, their surface migration, adsorption and desorption rates, and the activation energies of these processes can be measured in a wide temperature range. This is done by immersing the entire microscope into a cryogenic bath and by heating the emitter to any desired temperature up to near its melting point. Most studies have been done with the high-melting metals, W, Re, Ta, Mo, Nb, Ir, Pt, Rh, Pd, V, Ni, Fe, Ti, Cu, and various alloys and with adsorption layers of H_2 , N_2 , O_2 , CO, CO_2 , and the noble gases. Detailed patterns of individual organic molecules, such as phthalocyanine or smaller aromatic compounds, cannot yet be fully explained. The interpretation of thermal desorption experiments is often difficult because of the occurrence of surface migration, while the well-defined field strength obtained in field desorption experiments cannot be fully utilized because of the complex polarization conditions at the various adsorption sites.

The field ion microscope (Fig. 2) was developed to overcome the limitation of resolution of the field electron microscope. In contrast to the case

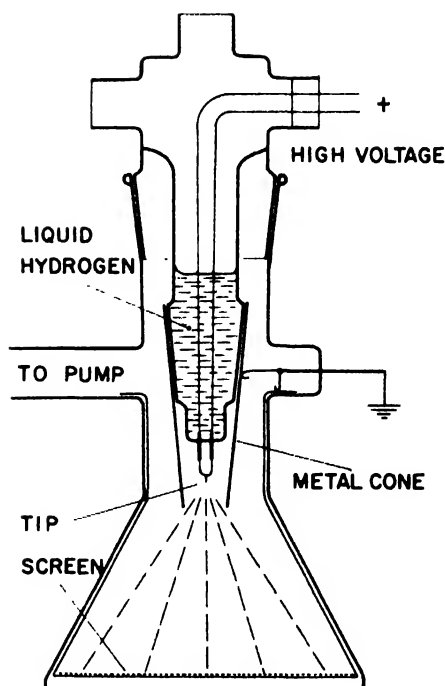


FIG. 2. Schematic diagram of a demountable field ion microscope. The hemispherical cap of the tip, having a typical radius of 500 \AA , is radially projected onto the fluorescent screen with the help of helium ions.

of electrons, the undesirable tangential motion of the imaging ions can be reduced by lowering the emitter temperature. The shorter de Broglie wavelength of ions is also an advantage. Usually operated with helium as the imaging gas and the tip cooled by liquid nitrogen, or better by liquid hydrogen, the ions image the field distribution over the emitter tip in a resolution of up to 2.5 \AA (Fig. 3). The specimen is atomically cleaned and shaped to an evenly curved, highly perfect surface by field evaporation. Helium atoms ionize above each atomic protrusion of the specimen and are then accelerated towards the screen. Each protruding surface atom is imaged by about 10^5 helium ions/sec. The surface stays atomically clean as all impurity gases have a lower ionization potential than helium and thus are ionized in the lower field region in space before they can reach the tip. Vacuum requirements are, therefore, very modest, and an unbaked demountable system can be used. Image stability requires that the evaporation field be above the ionization field of the imaging gas. Field ion microscopy with helium is, therefore, limited to the refractory metals. Using lower ionization potential gases, such as Ne, Ar, or H_2 , imaging with some loss of resolution can be done with reduced fields, and the common transition metals become accessible. Extremely low image intensity and the instability of the surface near the evaporation field can be partly overcome with the use of photoelectronic image intensification. The

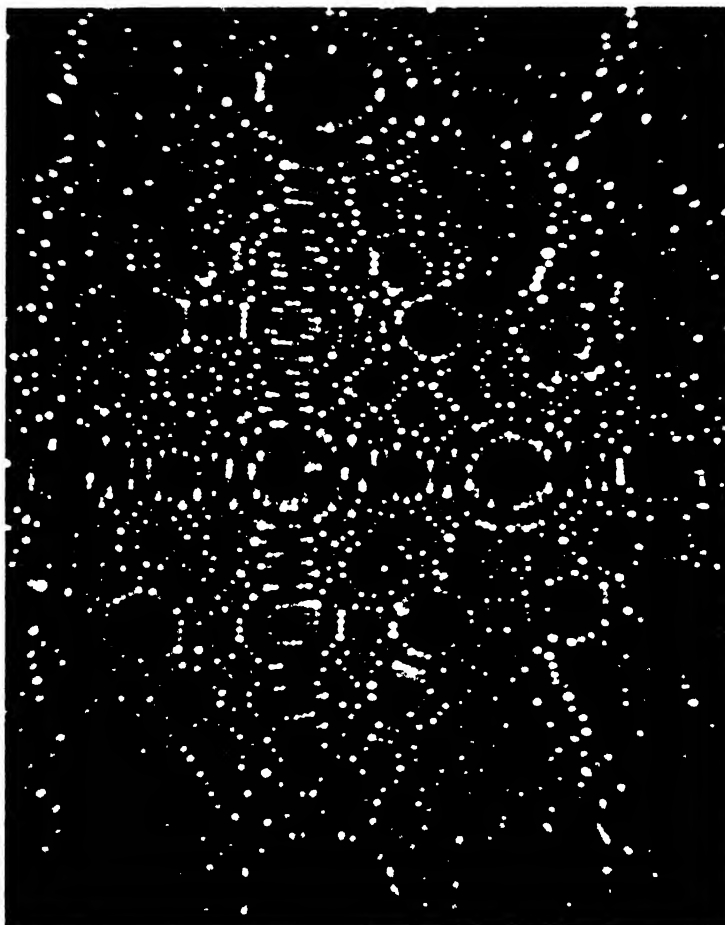


FIG. 3. Field ion micrograph of a platinum-cobalt alloy crystal (50 per cent, ordered at 700 C), showing the individual atoms as single dots. Original magnification on the 5-inch screen was 1.5 million diameters.

most promising application is the direct observation of lattice imperfections in metal crystals. Single vacancies produced by quenching, cold working, or irradiation can be counted. Interstitials resulting from radiation damage can be seen; their thermal migration to the tip surface can be measured. The core structure of dislocations and the match of the lattices along grain boundaries can be seen in atomic details, including their interaction with impurity atoms. Controlled field evaporation permits the removal and inspection of subsequent surface layers, so that the depth of the crystal can be searched for defects. The high mechanical stress $F^2/8\pi$ of the field, amounting to nearly 1 ton/mm² at the helium imaging field, can influence the defect structure of the specimen, and it can also be utilized for the study of slip bands and fatigue. To a limited extent, short-range and long-range order in alloys can be studied, although it is difficult to distinguish between the constituents. The first stages of corrosion by adsorbed gases and the

polarization modified adsorption sites of individually visible gas molecules are other promising subjects of field ion microscopy.

F. W. MULLER

References

- Good, R. H., and Muller, F. W., "Field Emission," in "Handbuch der Physik," (Encyclopedia of Physics), Second edition, Vol. XXI, pp. 176-231, Berlin, Springer, 1956.
- Dyke, W. P., and Dolan, W. W., "Field Emission," *Advan. Electron. Electron Physics*, **3**, 90-182 (1956).
- Muller, E. W., "Field Ionization and Field Ion Microscopy," *Advan. Electron. Electron Phys.*, **13**, 83-179 (1960).
- Gomer, R., "Field Emission and Field Ionization," Cambridge, Harvard University Press, 1961.

Cross-references: ADSORPTION AND ABSORPTION, ELECTRON MICROSCOPE, POTENTIAL, SOLID-STATE PHYSICS, SOLID-STATE THEORY.

FIELD THEORY

The description of the physical world has evolved profoundly through the ages, at times because of, at other times being the cause of, sweeping changes in our philosophical, mathematical and experimental knowledge. Greek geometry concerned itself essentially with properties of "objects as such," a triangle or a cube being studied, for example, without any thought of their spatial environment; the Ptolemaic system enhanced this view into a clockmaker's dream, where celestial bodies parade around the earth, rigidly driven in circular motions. Only with Descartes' analytical geometry did objects become "portions of space" and the properties of space itself the main object of study; with Galileo and Newton, a correct science of dynamics was born, which permits the prevision of an amazing number of mechanical phenomena in that space from a few first principles.

Field theory studies the phenomena of the physical world as due to interactions which propagate through space; the "geometrical emptiness," which is the space of mathematics, becomes the medium into and through which actions take place or, even more drastically, a structure which is itself determined by the properties of matter, as in general relativity (which will not be discussed in this article).

Suppose two bodies interact in space, e.g., the sun and the earth with Newton's law, or two electric charges with Coulomb's law. Two pictures of this situation are equally possible and correct. One is that this interaction cannot be conceived if *both* bodies are not there and that we should study primarily its effects without looking for a detailed mechanism for its propagation from one body to the other; this is the description of "action at a distance", in which *forces* are the main concepts and space is a vacuum into which bodies follow trajectories determined by the forces acting upon them. The other picture consists in imagining that each body, whether alone or not, modifies the structure of the space which surrounds it, geometrically or because in each point of that space there is now potentially a force, which becomes active if another body occupies that point, but should be conceived as existing there in any case; the main objective is here to study how these "fields of force" are created in space by material objects and how they propagate; this is the point of view of "action with contact," which finds its full development in field theory.

Mathematically, a field is characterized by assigning to each point of space a quantity which is *intrinsically* associated with it; a temperature, for instance, or a velocity, or a tensor or a spinor of arbitrary rank. "Intrinsic" means that if we change our frame of observation, this quantity does *not* change; supposing, e.g., that our field is that of the velocities at a given instant of all the points of a moving fluid, if we rotate our coordinate system we shall observe different values for the components of those velocities, just

because we, not the velocities, have changed position. It is therefore essential that, together with the specification of the field quantities, their transformation laws also be assigned under changes of the reference frame; these laws are indicated by the description of the field quantity as a "scalar" (which does not change), a "vector" (which changes with the same law as the coordinates), a "tensor," etc.; the complete specification of all such possible laws is a standard chapter of group theory.

Physically, we have to account for the creation or the existence of the field, by describing the field quantities as generated by "sources," such as positive or negative charges for the electromagnetic field, or the sources and sinks of hydrodynamics. Moreover, we have to describe in which way the values of the field quantities change when the point at which they are considered, or the time, is changed. In the absence of discontinuities, for instance in vacuum, one expects these values to differ by infinitesimal amounts if the corresponding points are infinitesimally close, in some way which is typical of the field considered; in other words, that the rates of change of the field quantities with respect to the space coordinates and time be connected by relations which specify both how these changes can occur compatibly with the geometrical properties of space, and how they are related to the sources. Group theory determines all the possible forms which are permissible for these relations, which take the name of *field equations*; each field theory is characterized by a special set of field equations, which are clearly *partial differential equations*.

RELATIVITY and QUANTUM THEORY have played a great rôle in the development of field theory; we shall briefly discuss, later, their influence both in the explanation of new physical phenomena and in the mathematical formulation of the theory.

Field theory has taken an entirely new shape with the so-called second quantization, which has led to several modern developments, of which some embody faithfully the concepts outlined thus far and others instead represent new views in natural philosophy; this is still a matter of controversy at present, and it is yet unpredictable whether a reasonably lasting description of nature will come out of such attempts or whether a new drastic turn in human thought will be necessary before we can hope to understand the fundamental laws of physics. Be that as it may, the ideas and the computational techniques of field theory have proved already of invaluable help in the description of many phenomena, from particle physics to superconductivity.

It is convenient to examine first the theories in which the field quantities are ordinary functions of space and time points, regardless of whether they have a direct physical meaning (as with the velocities of hydrodynamics and the electromagnetic forces) or not (as with the wave function which obeys a Schrödinger equation). This comprises of course most of classical and modern physics; mathematics permits again,

however, a tremendous conceptual simplification. In the study of continuous media or fields one is, most often, interested in one of the following classes of phenomena:

(1) Phenomena which consist of the propagation of some action; the medium itself is not transported from one place to another; typical is the propagation of waves, whether they be seismic, fluid or electromagnetic;

(2) Phenomena in which there is transport or diffusion of a quantity in a medium: of heat in a wall, of solute in a solvent, of neutrons in a pile;

(3) Equilibrium phenomena: deformations of strained elastic bodies, electro- or magnetostatic fields as determined by charges and boundaries.

Each class is ruled by essentially one type of equation. Let $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ denote the Laplace operator; $\phi = \phi(x, y, z, t)$ the field quantity; F some function of x, y, z, t, ϕ and, at most, of the first-order derivatives of ϕ ; v a velocity; and D a diffusion constant. The corresponding equations can be brought into the standard forms:

$$\Delta\phi - \frac{1}{v^2} \frac{\partial^2 \phi}{\partial t^2} = F \text{ (hyperbolic partial differential equation) (1)}$$

$$\Delta\phi - \frac{1}{D} \frac{\partial \phi}{\partial t} = F \text{ (parabolic partial differential equation) (2)}$$

$$\Delta\phi = F \text{ (elliptic partial differential equation) (3)}$$

If the field quantity has more than one component, one may deduce for each of its components an equation which is essentially of the same type, although it may be difficult or impossible to obtain an independent equation for each component.

This classification of physical phenomena according to the type of equation to which their study can be reduced is of the greatest importance: Eqs. (1), (2) and (3) are called in fact "the equations of mathematical physics"; more specifically, Eq. (1) is also called the wave equation, Eq. (2) the heat equation, and Eq. (3) the Laplace or potential equation. The study of their mathematical properties gives complete information on all the physical phenomena which they describe.

The equations of quantum mechanics can also be brought, at least formally, into the form of Eq. (1) or (2); the intervention of complex quantities modifies the situation somewhat, in a way which we cannot discuss here.

Electromagnetic phenomena fall typically into the category of Eq. (1): each of the components of the electric field $\mathbf{E} \equiv (E_x, E_y, E_z)$ and of the magnetic field $\mathbf{H} \equiv (H_x, H_y, H_z)$ satisfies, in vacuum, Eq. (1), with $F = 0$; the connections between \mathbf{E} and \mathbf{H} are given by the Maxwell equations, which characterize completely the theory, and lead in vacuum to the result just

mentioned. When \mathbf{E} or \mathbf{H} does not vary with time, Eq. (1) reduces to Eq. (3), thus yielding electro- or magnetostatics.

The electromagnetic field, i.e., the vectors \mathbf{E} and \mathbf{H} , generated by a distribution of moving charges or currents confined within a limited volume has a part which becomes dominant at a large distance from that volume, because it decreases only with the inverse of that distance (instead of the inverse-square law of static fields); this part constitutes the *radiation* field, which is responsible for the transmission of energy and signals (the radiated energy is, of course, supplied by the mechanism which drives the generating charges or currents). This is easy to understand: the energy radiated through a large sphere around the source is proportional to the area of the sphere times the square of E ; it vanishes therefore with increasing radius for all but the radiative component, for which it stays constant: energy is actually removed from the source and radiated away to all distances. The study of radiation is a most important part of the theory, both macroscopically (telecommunications, radar) and microscopically (atoms, nuclei, elementary particles).

Relativity and quantum mechanics have extended and modified profoundly the classical picture presented so far. The very concepts of space and time change with special relativity: events which are simultaneous for an observer are not such when seen by another observer in uniform motion with respect to the first, because time and space are mixed together by the Lorentz transformations which relate the reference frames associated with the two observers. As a consequence, the laws of nature can retain their universal validity only if they are formulated in the same form by any such observer, i.e., if their form is not altered by a Lorentz transformation—technically speaking, if they are "Lorentz covariant." This requirement becomes a stringent dogma; it suffices to determine, with the help of group theory, the possible equations for any conceivable relativistic field theory; it is of invaluable help, when computations are made, in checking or correcting them.

The nonrelativistic Schrödinger equation for the wave function of a particle is, but for the appearance of complex quantities, of the type (2) described before: this is not acceptable in a relativistic world, because time and space are not treated alike. One needs either an equation which contains only second-order derivatives, or one with only first-order derivatives; for a free particle, this leads either to the Klein-Gordon equation, which is of type (1), or to the Dirac equation, which contains linearly only the first-order derivatives of the wave function, but has a mathematical structure which necessarily assigns special physical properties to the particles described by it. It was one of the greatest triumphs theoretical physics ever witnessed, to discover that such properties are actually displayed by all particles which obey the Dirac equation: spin, and the existence for each Dirac particle of a

corresponding *antiparticle*, i.e., a particle having the opposite mechanical and electrical properties.

The requirement of relativistic covariance has thus led to fundamental physical discoveries; for each particle obeying the Dirac equation, the corresponding antiparticle has been experimentally found in nature; electron and positron, proton and antiproton, neutron and antineutron, etc. What is more, the theory predicts that a particle-antiparticle pair can be created in a collision phenomenon, if sufficient energy is available, or can annihilate itself, giving away its energy in the form of electromagnetic radiation or other particles.

The classical theory allowed only for the electromagnetic radiation emitted by moving charges or currents; the creation or absorption of particles in collision phenomena, as well as the creation or annihilation of pairs, were outside its scope and possibilities. A new formulation of the theory was needed, which could account consistently for all such phenomena, handling situations in which particles can be created and destroyed in any numbers. The formalism devised for this purpose is that of quantum field theory.

The basic idea is to describe each type of particle by means of a field which is not any more an ordinary numerical function of space and time, but an "operator," i.e., a quantity which changes the number of particles existing in any given state of the system. If the field operator is known, one can then evaluate the probability of a given state (so many particles, with determined energies and momenta) changing into an equally determined, different state. If the particles do not interact among themselves or with other particles, no change is possible; if there is interaction, the field operator has a structure which can cause such transitions. The field equations appear to be essentially the same as those of the classical Maxwell, Klein-Gordon, Dirac theories, etc.; their structure is however fundamentally different, because they now must be equivalent to infinite sets of ordinary equations, which couple states with different and ever-increasing numbers of particles.

Fields which are associated with particles obeying Bose-Einstein statistics (of which any number can be found in any given state) have radically different mathematical properties from fields associated with particles obeying Fermi-Dirac statistics (of which at most one can be found in any given state); examples of the first are photons (the massless neutral quanta of the electromagnetic field), pions (massive particles, with or without electric charge, which are believed to be responsible for nuclear forces), etc.; examples of the second are electrons and positrons, protons and antiprotons, etc.

The passage from numerical fields to operator fields is called "second quantization"; quantum field theory deals with operator fields.

Striking successes have been met with this approach. From a quantitative point of view, they are confined mostly to electrodynamics,

where very small deviations from the values predicted by the non-quantized theory, which were observed in the measurement of the magnetic moment of the electron and in the so-called Lamb shift, were accounted for with amazing accuracy by quantum field theory. Qualitatively, the new conceptual framework has proved extremely useful in understanding elementary phenomena, especially with the help of the diagrams devised by R. P. Feynman, which give a simple intuitive picture of collision and radiation processes involving elementary particles. Very little has been achieved quantitatively, though, for theories other than electrodynamics, because of the tremendous mathematical difficulties which arise as soon as the simplest approximation techniques are not applicable because the interaction is too strong; nevertheless, these ideas have proved greatly helpful in many ways, in combination with general principles of symmetry, Lorentz invariance and causality.

A beautiful consequence of this conception, which assumes that particles are the quanta of a field (as photons were recognized by Einstein to be the quanta of the electromagnetic field) was the discovery of H. Yukawa, that whenever such quanta have a mass different from zero, the force they create between two bodies which interact by exchanging such quanta with each other must be an exponentially decreasing function of distance; this force can become of a coulombian type only if the mass of the quanta vanishes. Thus, the Coulomb force can be explained as due to the exchange of photons among electric charges, the nuclear forces (which have typically short ranges) as due to the exchange of massive particles among nucleons. Exchanges of this nature are not observable in the laboratory, because this runs against Heisenberg's indeterminacy principle; if enough energy is supplied, however, such quanta can actually break loose and do appear as the particles created in collision processes.

The mathematical difficulties encountered in quantum field theory are many, and there is as yet lack of agreement as to the best way to circumvent some of them. Besides mathematical complexity, which prevents all but the simplest calculations, there are many unsolved problems of mathematical rigor and apparent inconsistencies which can be removed only by delicate analyses. Typical of the latter is the fact that unsophisticated calculations give infinite values for masses and charges of interacting particles, and a painstaking analysis is required to retrieve from them the significant physical values; this is the so-called renormalization procedure, which copes with infinities which partly are already present in the classical theory (such as the infinite electromagnetic contribution to the mass of a point-like charged particle, when computed from Maxwell's equations) and partly originate from the new formalism (which permits, for instance, pair creation).

It is not yet certain whether such difficulties are due to the lack of adequate mathematical techniques or are the expression of a fundamental

inadequacy of the theory to describe ultimate laws of nature. For this reason, while, on the one hand, the attention of some theoreticians has been directed to perfecting the mathematical foundations of quantum field theory (giving rise to axiomatic field theory and to more rigorous methods of obtaining and studying the quantum field equations, etc.), on the other hand, most physicists have been trying new avenues, such as the S-matrix theory, dispersion relations, the so-called Regge poles, etc.; these approaches have certainly led to very useful results, but they leave altogether at least as many doubts as hopes.

Whatever may be the future prospects of field theory as the correct means for describing the fundamental laws of nature, its tremendous usefulness in providing a conceptual framework, in inspiring new ideas, and in suggesting computational techniques has been overwhelmingly demonstrated in the last decades. It has now found a new, very fertile ground of application in the study of systems containing a very large number of particles, where it has already provided a reasonably good quantitative understanding of superconductivity and superfluidity, and promises many other results of interest in the study of solids and liquids.

E. R. CAIANIELLO

Cross-references: ELEMENTARY PARTICLES, QUANTUM THEORY, RELATIVITY, SCHRÖDINGER EQUATION, VECTOR PHYSICS

FISSION*

Nuclear fission is the breakup of a heavy nucleus, such as that of uranium, into two medium-weight nuclei, with the release of a considerable quantity of energy. Also produced are a few neutrons, some gamma rays, and a number of beta-particles (electrons) from the radioactive decay of the two fragments. Fission occurs spontaneously in some cases, or may be induced by bombardment of the fissionable material with neutrons, protons, or other particles.

Discovery of Fission. Although fission was not discovered until 1939, it had been realized, ever since Einstein published his theory of relativity in 1905, that there was a theoretical possibility of releasing tremendous energy from matter.

Fission is now known to have been first produced by Enrico Fermi and his co-workers in 1934, when they irradiated many elements, including uranium, with the newly discovered neutrons. They found a number of different β -activities to be produced from uranium, but believed that these were due to neutron capture. Later radiochemical work indicated that some of the new activities were from elements chemically similar to the much lighter elements Ba, La, etc.

Fission remained unrecognized until O. Hahn and F. Strassmann, German radiochemists,

showed by very careful work that these products were not merely chemically similar to lighter elements, but *were* lighter elements. They published their startling results in the January 6, 1939 issue of *Naturwissenschaften*. On January 16, Lise Meitner and O. Frisch sent in to *Nature* (from Stockholm and Copenhagen) a paper in which they named the new process "fission," predicted that the fragments should have large kinetic energies, and explained the process in terms of a liquid-drop model. On the same date, Frisch sent in to *Nature* another paper in which he reported that he had observed the large electrical pulses from fission fragments in an ionization chamber.

Niels Bohr in the meantime had taken the news of the discovery of fission, and the prediction of large energies, to a conference on physics in Washington, D. C. During the conference physicists in a number of laboratories independently verified the tremendous kinetic energies of fission fragments, unaware as yet of Frisch's results.

During the months following the discovery of fission, there was feverish activity in laboratories around the world. It was soon discovered that neutrons were being produced by fission, and that almost all of the fission of U was taking place in the relatively rare isotope, U^{235} . In that same year (1939), Bohr and Wheeler published their theory of fission, based on the liquid-drop model, which is still basic to modern fission theory.

Development of Atomic Energy. On the date of publication of the Bohr-Wheeler paper, September 1, 1939, Germany invaded Poland, the Second World War was underway, and fission suddenly had a new importance. It was realized by many that a chain reaction was possible for fission, with the neutrons from each fission producing more fissions, resulting in the release of very large amounts of energy.

The fission process results in the conversion of 0.09 per cent of the mass of the original nucleus into kinetic energy. This amounts to about 200 MeV per fission, or 3.20×10^{-11} joules. The fission of 1 kilogram of U^{235} thus releases a total energy equal to 8.21×10^{14} joules, or 2.28 $\times 10^7$ kilowatt-hours. This is roughly equal to the daily output of Hoover Dam, and very much greater than the energy released in chemical reactions. One kg of U^{235} is equivalent in energy release to the burning of 3.45×10^6 kg of coal (C) by 9.20×10^6 kg of oxygen. This chemical process releases 7.2×10^{11} of the mass as heat energy, more than 10 million times smaller than for fission.

The fission of 1 kg of U^{235} is also equivalent in energy released to the detonation of 19.6×10^3 kg (19.6 kilotons) of high explosive. A kiloton, 1000 metric tons, is conventionally taken to be equivalent to 10^{12} calories, or 4.186×10^{12} joules; a megaton is 1000 times larger.

Thus it was known to many people in 1939 that it might well be possible to produce the destructive effect of many thousands of tons of high explosive with a single bomb containing a

* Work performed under the auspices of the U.S. Atomic Energy Commission.

relatively small amount of fissionable material. It seemed probable that Germany would press ahead with this development. Aware of this danger, scientists in the rest of the world largely ceased publishing fission results by 1940.

Work on fission was continued quietly at an increasing rate. In June 1942, the Manhattan Project (under U.S. Army direction) was set underway in the United States, with the objective of producing nuclear weapons, if possible. The first chain reaction was produced on December 2, 1942, under the direction of Enrico Fermi, who had arrived from Fascist Italy in January 1939. Fermi and his co-workers had piled up blocks of ordinary uranium and extra-pure graphite (carbon) to produce a nuclear reactor, under Stagg Field Stadium of the University of Chicago. The carbon was used to slow down fission neutrons and thus increase the likelihood of fission. Cadmium rods inserted in the reactor (at that time called a "pile") were used to control the chain reaction by capturing a certain fraction of the neutrons.

Thus, by December 1942, a fission chain reaction had been achieved. In the following years, research vital to the Manhattan Project was carried on at many laboratories. In order to produce the fissionable material for nuclear weapons, two tremendous industrial plants were set up beginning in 1943, at Oak Ridge, Tennessee, and Hanford, Washington.

The Oak Ridge plant was for the purpose of separating the more fissionable isotope, U^{235} , from the much more common isotope, U^{238} , which accounted for 99.3 per cent of the mass of ordinary uranium metal. The most successful separation process, which is still in use at Oak Ridge, was that of gaseous diffusion. Uranium in the form of a gas, uranium hexafluoride, is passed through a long series of porous barriers. The lighter isotope U^{235} can diffuse more readily than U^{238} , and the result of the process is enriched uranium, the U^{235} content having been increased from 1 part in 140 to perhaps 95 per cent.

The Hanford plant consists of giant nuclear reactors to produce a new element, plutonium, from ordinary uranium by neutron capture. It was thought, and eventually proved, that plutonium should be fissionable by slow neutrons in the same way as U^{235} . Capture of a neutron by U^{238} produces the heavier isotope U^{239} , which then decays by beta-emission to the new element Np^{239} (neptunium), and then by another beta-decay to Pu^{239} .

The work of designing and building nuclear weapons was carried out at a laboratory set up at Los Alamos, New Mexico. This laboratory, under the direction of J. Robert Oppenheimer, began its work in early 1943. Before the end of the war, a good fraction of the world's most eminent nuclear physicists had come to work at Los Alamos, including Bohr, Fermi, Frisch, and many British scientists.

One of their most basic problems was to find a way to assemble the fissionable components of a

weapon rapidly enough to produce a powerful chain reaction, lasting less than one μ sec. The individual masses of enriched uranium or plutonium had to be of such a size and shape as not to be capable of a chain reaction; i.e., they had to be of less than critical mass. If these pieces were not assembled at sufficient speed, the result would be only a minor explosion, or perhaps the melting of the device. Two approaches were tried. One involved firing one piece of fissionable material at another in a short "gun." The other was the implosion method, in which the fissionable material is assembled into a highly compressed mass by the explosion of a surrounding spherical shell of high explosive.

Both methods of achieving a nuclear explosion were ultimately successful, but it was not known whether either method would work until July 16, 1945, when the first nuclear weapon ("atomic bomb") was set off in a desert area near Alamogordo, New Mexico. On August 6 a weapon was exploded over Hiroshima, Japan, and on August 9 another was detonated over Nagasaki. Each weapon had a yield equivalent to about 20 kilotons of high explosive and caused tremendous destruction. On August 10, the Japanese first offered to surrender, and accepted Allied terms on August 15; the mobilization of the massive United States invasion force was called off. Germany, which had surrendered on April 8, 1945, was found to have made little progress toward nuclear weapons during the war.

Following the end of the war, the work on nuclear weapons and nuclear power in the United States were placed under the newly created (1946) Atomic Energy Commission. Similar agencies have since been set up in many countries. On August 29, 1949, the U.S.S.R. detonated its first nuclear device. During the next several years, enormous production plants for U^{235} were built at Paducah, Kentucky, and Portsmouth, Ohio, as well as facilities for producing Pu^{239} and hydrogen isotopes on the Savannah River in South Carolina. On November 1, 1952, the Los Alamos Scientific Laboratory exploded the first thermonuclear device ("hydrogen bomb"), with a yield equivalent to many megatons of high explosive. Such a weapon uses nuclear fission to trigger nuclear fusion, in which light elements (such as the various isotopes of hydrogen) are combined at exceedingly high temperature to give heavier nuclei, with considerable release of energy. On August 12, 1953, the Russians set off their first thermonuclear weapon. In the years since then, the British, French, and Chinese have all developed their own nuclear weapons.

Nuclear reactors and nuclear power have been developed extensively since the war, in many countries. Nuclear-powered submarines, aircraft carriers, and other vessels have revolutionized naval strategy. Nuclear-propelled rockets (the "Rover" program at Los Alamos) will probably soon be used in space exploration. The use of nuclear explosives for earth-moving and other engineering purposes has been an extensive project ("Plowshare") of the Lawrence Radiation

Laboratory at Livermore. Nuclear power plants are rapidly becoming more widespread, and more economical.

Chain Reactions. The basic feature which makes both nuclear reactors and nuclear weapons possible is the "chain reaction," in which each fission produces another fission, or several, by means of the several neutrons emitted from fission. If there is not enough fissionable material, or it is not arranged compactly enough, no chain reaction will be possible. The fission neutrons emitted in such a situation will have too great a chance of escaping from the fissionable material, or of being absorbed in non-fissionable material, to continue the chain of fissions should one fission occur.

There is thus a "critical mass," or minimum amount of fissionable material, necessary for a chain reaction in any given arrangement. For a spherical mass of metal in air, the critical mass of highly enriched (94 per cent) U^{235} , for instance, is 52 kg; the critical mass for U^{233} or Pu^{239} metal is lower, about 16 kg. The critical mass can be lowered by mixing fissionable material with graphite or other material as a "moderator" to slow down the neutrons, or by surrounding the fissionable material with a reflector to scatter neutrons back.

The smallest critical masses are achieved by water (or heavy water) solutions of fissionable material, since H and D are most effective in slowing down neutrons. The "Water-Boiler" thermal reactor at Los Alamos operates ordinarily with about 800 grams of enriched uranium in water solution, and has gone critical with less than 600 grams.

The energy of the neutrons producing most of the fissions in a reactor determines whether it is called "fast," "intermediate," or "thermal." Fast reactors use neutrons only slightly slowed down from the 2-MeV average energy with which they are emitted. Thermal reactors use neutrons slowed down, by collision with moderator atoms, almost to the velocities corresponding to thermal motion at the temperature of the reactor. At 294 K, about room temperature, the most probable velocity is 2200 meters/sec., which corresponds to a neutron energy of 0.0253 eV. Intermediate reactors use partially slowed down neutrons, with less than 100-keV energy.

In order for a chain reaction to continue, it is necessary that each fission produce, on the average, at least one more fission. The average number of fissions produced by the neutrons from one fission is called k , the criticality factor or multiplication constant. If k is less than 1.0 (the critical value), a chain reaction, even if begun with many fissions, will soon die out.

A reactor must have $k = 1.0$ during normal, steady, operation, and must have k greater than 1.0 during the start-up operation. These changes in criticality factor can be brought about by introducing or removing from the reactor control rods made of neutron-absorbing elements, such as cadmium or boron.

This regulation of reactor power is made much

easier by the phenomenon of delayed neutrons, which are emitted from some fission products for a few seconds after the fission. In the case of thermal-neutron fission of U^{235} , 0.65 per cent of the neutrons produced (0.0158 out of 2.42 neutrons per fission) are delayed in this way. Only if k were increased above 1.0065 in this case could the reactor power increase rapidly, as each fission would then, on the average, produce one more fission promptly. Reactors are of course designed to avoid this "prompt critical" condition, as this could lead to overheating and destruction of the reactor. In any case, reactors cannot possibly achieve the high values of the criticality factor needed for a nuclear explosion.

The Fission Process. The fission of a nucleus can be understood on the basis of the liquid-drop model, which was first discussed by Meitner and Frisch and later greatly extended by Bohr and Wheeler. This model explains many features of fission, but others are still not fully understood, and the complete theory of fission is still awaited.

On the basis of this model, the nucleus is assumed to be similar to a uniformly charged drop of incompressible liquid. It will then be normally spherical, kept in this shape of minimum energy by the effect of surface tension. However, the individual parts of the positive charge, actually protons, tend to repel each other and to lessen the effective surface tension. It has been calculated that the two effects will cancel each other out for a value of Z^2/A equal to about 45 (Z is the atomic number or the number of protons; A is the mass number or total number of nucleons in the nucleus). The ratio of Z^2/A to this critical value is called the fissionability parameter x .

For a fissionability parameter equal to 1.0, the nucleus should have no effective surface tension and no stability against distortion, so that it should promptly elongate to a point where the coulomb (electrostatic) forces can blow it apart. Such a nucleus would have no "fission barrier" and could not exist long. Known nuclides have lower values of the fissionability parameter; for U^{235} the value of x is about 0.8. For such nuclei, it takes 5 or 6 MeV of energy to deform the nucleus to the point where coulomb forces can cause fission. However, such a fission barrier may be overcome by the capture of a neutron, which adds an excitation energy of 5 or 6 MeV.

For the even- Z , odd- N nuclides U^{233} , U^{235} , and Pu^{239} (N is the number of neutrons in the nucleus), fission can be induced by the capture of even a low-energy neutron. Even-even target nuclides require more energy, so that U^{238} is very unlikely to fission upon capture of a thermal neutron, but needs about 1 MeV additional neutron energy to overcome the fission barrier. Even-even compound nuclei (i.e., after neutron capture) are evidently more likely to undergo fission. Even-even nuclei are also much more likely to undergo spontaneous fission, in which an occasional nucleus manages to overcome the fission barrier without any added energy. Fission may also be induced by gamma

rays of energy equal to the fission barrier or by energetic particles of many kinds such as protons or alpha particles.

The probability that particles passing through a target of fissionable material will induce fission is measured by the fission cross section. The "cross section" is the effective area of a nucleus for a given process. For a thin target of thickness t / cm having n nuclei per cubic centimeter, each with fission cross section σ cm², the probability that a particle passing through the target will induce a fission is $n\sigma t$.

Figure 1 shows the fission cross section for 11 fissionable nuclei, as a function of the energy (in MeV) of the neutron inducing fission. For those nuclides (U^{233} , U^{235} , Pu^{239}) which lead to even-even compound nuclei, the cross sections are very high for low-energy neutrons; these are "fissile" nuclides. For the other target nuclides, the fission cross section is extremely small until the neutrons have energy in excess of a threshold energy. All the cross sections may be seen to increase noticeably at about 7 MeV. This increase is due to the added possibility of fission following low-energy neutron emission, which becomes energetically possible here.

The energy released when fission takes place amounts to about 200 MeV. This is the difference in mass between a heavy fissionable nucleus and the two medium-weight nuclei, plus neutrons, into which it breaks up. This energy release of 200 MeV amounts to 21.3 per cent of the mass of a single proton or neutron. Most of the energy

appears in the form of kinetic energy of the two fission fragments, which are sent flying apart with an average total of 167 MeV in the case of thermal fission of U^{235} .

The two fragments share the kinetic energy in the inverse ratio of their masses, since they have equal and opposite momenta. The mass division is usually asymmetrical, as may be seen in Fig. 2. The probability that neutron fission of U^{235} will produce two given fragments of nearly equal mass (symmetrical fission) is about 600 times lower than that of the most probable case of asymmetrical division into light and heavy fragments. The initial fragments (before neutron emission) in this case have masses averaging 96 and 140 mass units. The lighter fragment has on the average about 99 MeV of kinetic energy, and the heavy fragment about 68 MeV. These energies correspond to initial velocities of 1.4 and 1.0×10^9 cm/sec, or 4.7 and 3.2 per cent of the speed of light, respectively. The total fragment energy is rather closely given, on the average, by $0.121 Z^2/A^{1/3}$ MeV.

Such fragment energies are due to the Coulomb interaction between the two fragments, and would be expected on the basis of any fission theory. The asymmetrical mass division, however, would not be predicted by liquid-drop theory and has so far resisted detailed explanation. The liquid-drop model would, in its simplest form, predict primarily symmetrical fission, which is rarely observed.

The fragments, as we have seen, carry away

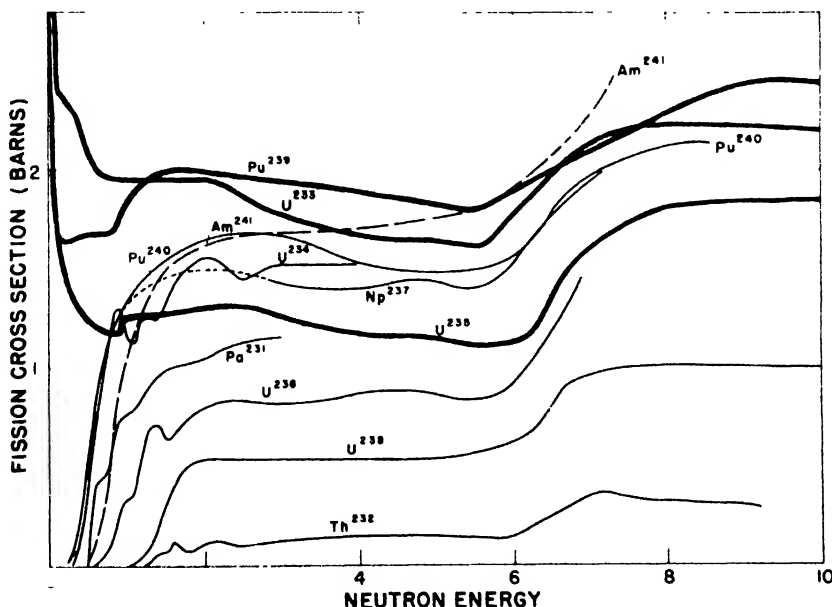


FIG. 1. Dependence of fission cross section (a "barn" is 10^{-24} cm²) on inducing neutron energy (in MeV) for 11 different fissionable nuclei (reprinted with permission from R. L. Henkel, "Fission by Fast Neutrons," in J. B. Marion and J. L. Fowler, Eds., "Fast Neutron Physics, Part II," New York, Interscience Publishers, 1963).

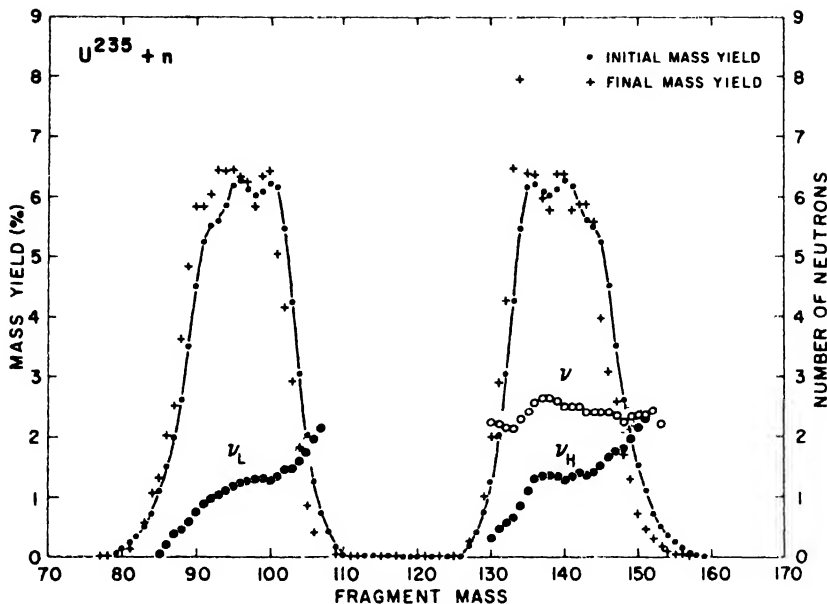


FIG. 2. Dependence of neutron yield on initial fragment mass for thermal-neutron fission of U^{235} . Average numbers of neutrons emitted by light and heavy fragments are given the symbols ν_L and ν_H , the total from both fragments is ν . Standard deviations are indicated by dotted lines. Also shown are the initial and final mass yields [reprinted from J. Terrell, *Phys. Rev.*, **127**, 880-904 (1962)].

most of the 200-MeV energy release in the form of kinetic energy. The rest of the energy is released in the form of neutron energy, gamma rays, and beta decay. The prompt neutrons, which are emitted within 10^{-11} sec. following fission, will be discussed below. The fission fragments are strong sources of prompt gamma rays, emitting about 8 within a microsecond or less following fission. The gamma rays have a broad spectrum of energies, up to as much as 7 MeV, but they average about 1 MeV apiece, so that the total energy emitted in prompt gamma rays is about 8 MeV. This is more than simple theories would predict.

Immediately after the emission of neutrons and prompt gamma rays, the fragments (now called "fission products") begin the process of beta decay, which ultimately accounts for perhaps 22 MeV of fission-energy release. The fission products have neutron-to-proton ratios which are nearly the same as that of the heavy nucleus from which they were formed (about 1.55). This is too many neutrons for stability in the fission-product mass region, where nonradioactive nuclei have neutron-proton ratios in the range 1.3 to 1.4. Since there is thus energy to be released by changing neutrons to protons, the result is a long sequence of beta decays averaging 3 or 4 for each fission product. Each beta decay results in the emission of a negative electron, a neutrino, and usually gamma rays, and an increase of the charge of the nucleus by one unit. The neutrinos ultimately carry away about 10 MeV of the energy released, and for all

practical purposes, this energy is not detectable and is never seen again. The fission products are intensely radioactive immediately after fission, as the beta decays having the shortest half-life are completed, and gradually become less radioactive with the passage of time.

For a small fraction of the fission products it is energetically possible for a neutron to be emitted at some stage in the chain of beta decays. These delayed neutrons, amounting on the average to 0.0158 per thermal-neutron fission of U^{235} , are very useful in stabilizing the operation of nuclear reactors, as mentioned above. Because they are emitted immediately following beta decays, they appear to follow the same radioactive decay curves as some of the beta decays, and have varying half-lives, of the order of seconds.

In rarer cases - about one fission out of 400 - an alpha particle (He^4) is emitted during the fission process, in addition to the two fragments. This process is usually called ternary fission, as distinguished from the usual binary fission. Because of the high energy (averaging 15 MeV) and direction (roughly perpendicular to fragment motion) of the alpha particle, it is probably formed at the same time as the two heavier fragments, and between them. In still rarer cases, a lighter charged particle such as a nucleus of tritium (H^3) is seen. There seems even to be evidence that, in perhaps one fission out of 100,000, three fragments of roughly equal mass are formed.

In the vast majority of fissions, however, two fragments and several neutrons are the result. For low-energy fission it is clear that most, or perhaps all, of the neutrons are emitted by the fragments. The average number of neutrons per fission ($\bar{\nu}$) ranges from 2.0 to 4.0 for various nuclides. Fission at high energies, such as that induced by 100-MeV alpha particles, produces many more neutrons, most of which were emitted before fission.

In the typical case of thermal-neutron fission of U^{235} , $\bar{\nu}$ is 2.42. For individual fissions the number ν can vary from zero to 5 or 6, the standard deviation from the average being ± 1.10 . The neutrons in this case have an average energy of 2.0 MeV, and the energy spectrum for U^{235} is often described by the Watt spectrum,

$$N(E) = -0.48394e^{-E} \sinh(\sqrt{2E})$$

Here $N(E)$ is the number of neutrons per unit energy E , in MeV, and the normalizing constant is chosen to give an integral of 1.0 over the entire spectrum. Another, more general, representation of the neutron spectrum is the Maxwellian distribution,

$$N(E) = (2/\sqrt{\pi T^3})^{1/2} E e^{-E/T}$$

in which T is a parameter equal to two-thirds of the average energy \bar{E} . Such a spectrum is predicted from nuclear temperature theory.

Of the roughly 2-MeV average energy per neutron, about 0.75 MeV is contributed by the motion of the emitting fragments. There is evidence that the average neutron energy increases with increasing number of neutrons. This would be expected from considerations of nuclear temperature, which lead to the relation $\bar{E} = 0.75 + 0.65 \sqrt{\bar{\nu} + 1}$, in MeV, in agreement with experimental data.

The total kinetic energy carried away by the typical 2 or 3 neutrons per fission is thus about 5 MeV, of which 2 MeV is taken from fragment energy. The fragments after neutron emission thus have a total energy reduced (in the case of U^{235}) from 167 to 165 MeV, and lower mass numbers.

The final fragment masses may be determined by radiochemical data on fission products. The final mass yields for U^{235} are shown in Fig. 2, as crosses. The initial mass yields may be determined from simultaneous measurement of the velocities of the two fragments and are also shown in the Figure. The differences between initial and final mass yields are accounted for by emission of neutrons and may be used to determine the numbers of neutrons. The average numbers of neutrons emitted by individual fragments are shown (ν_L and ν_H) as functions of fragment mass; the total $\nu (= \nu_L + \nu_H)$ is also shown.

As may be seen in Fig. 2, the average number of neutrons emitted by a fission fragment depends strongly on the fragment mass; it is near zero for the lightest of the light fragments, and also for

the lightest of the heavy fragments, and rises to high values elsewhere. Thus the two fragments from a fission will usually emit quite different numbers of neutrons, which implies quite different initial excitation energies. This phenomenon came as quite a surprise when it was first reported by Fraser and Milton in 1954. It has since been found to be a common and perhaps universal feature of fission.

It seems now that this "sawtooth" dependence of neutron number on fragment mass may be understood on the basis of varying stiffness of the fragments against deformation. The fragments having "magic" numbers of neutrons or protons would be expected to be stiff and to be exceptionally resistant to deformation and the consequent excitation. The magic number of neutrons, $N = 50$, will occur for fragments in the vicinity of mass 82, at the lower boundary of the light fragment peak. Similarly, two magic numbers ($Z = 50$ and $N = 82$) occur at the lower edge of the heavy fragment peak, near mass 130. The low neutron yields seen at these magic-number positions may be quantitatively explained on the basis of the effects of magic numbers on the deformation parameters. It is hoped that this new "fragment deformation" theory will be able to account quantitatively for the little-understood spectrum of fragment masses, since the mass yields and neutron yields tend to vanish near the same magic numbers.

JAMES TERRELL

References

- Glasstone, S., and Sesonske, A., "Nuclear Reactor Engineering," Princeton, D. Van Nostrand Co., Inc., 1963.
- Glasstone, S. Ed., "The Effects of Nuclear Weapons," revised edition, U.S.A.E.C., U.S. Government Printing Office, 1964; especially Chapter 1.
- Hyde, E. K., "The Nuclear Properties of the Heavy Elements III: Fission Phenomena," Englewood Cliffs, N.J., Prentice-Hall, Inc., 1964.
- Keepin, G. R., "Physics of Nuclear Kinetics," Reading, Mass., Addison-Wesley Publishing Co, 1965.
- Smyth, H. D., "Atomic Energy for Military Purposes," Princeton, N.J., Princeton University Press, 1946.
- Terrell, J., "Prompt Neutrons from Fission," in Proceedings of the I.A.E.A. Symposium on the Physics and Chemistry of Fission; I.A.E.A., Vienna (1965).
- Weinberg, A. M., and Wigner, E. P., "The Physical Theory of Neutron Chain Reactors," Chicago, University of Chicago Press, 1958.
- Wheeler, J. A., "Fission," in Condon, E. U., and Odishaw, H., Eds., "Handbook of Physics," New York, McGraw-Hill Book Co., 1958.

Cross-references: ATOMIC ENERGY, FUSION, NUCLEAR RADIATION, NUCLEAR STRUCTURE, NUCLEONICS.

FLIGHT PROPULSION FUNDAMENTALS

Propulsion is not a science in itself, and has no unique basic principles of its own. The fundamentals of propulsion really are selected from the basic laws of MECHANICS, THERMODYNAMICS, and CHEMISTRY; certain special types of propulsion require also some principles of ELECTRICITY or NUCLEONICS.

Propulsion is defined here as the act of changing the natural state of motion of a vehicle flying in air or in space. Thus a propulsion mechanism applies a force to a flying body and so changes the momentum of the body, or, in the case of maintaining steady, level equilibrium flight, the force of propulsion overcomes or equals the atmospheric drag. To propel a vehicle, it is necessary to accelerate another mass (usually a gaseous fluid) in the opposite direction. Hence, not only Newton's basic laws of motion, but also the conservation of momentum (change of momentum of vehicle equals change of momentum of ejected high-velocity masses) and the perfect gas laws are important here (see CONSERVATION LAWS AND SYMMETRY AND IMPULSE AND MOMENTUM).

The acceleration of a *gaseous working fluid* (often called propellant) to high exhaust velocities requires the supply and expenditure of energy. Thus, it is possible to classify flight propulsion systems according to their energy sources and the types of propellants (see Table 1). The device which creates and/or converts the energy into the form which is useful for propulsion is called an *engine*. It is the purpose of this section to summarize some of the most important propulsion relationships and to describe briefly the principal engine types.

Fundamental Relations. Assume an ideal engine inside of which the fluid receives energy and is heated and accelerated as shown schematically

in Fig. 1. The acceleration of a mass flow of air or fluid \dot{m}_a from an initial velocity v_0 (which equals the forward flight velocity of the vehicle) to the jet velocity at the exit v_e will result in a net thrust F , which is equal to the mass flow rate multiplied by the velocity increment. Additional

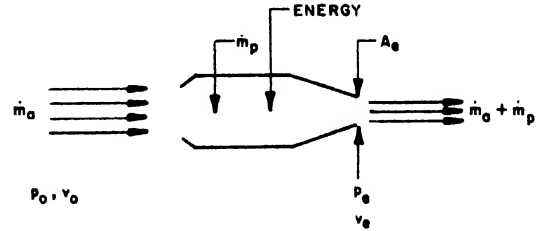


FIG. 1. Simple engine.

terms are added to this momentum relation in order to correct for the additional mass flow rate of the fuel or propellant (\dot{m}_p) (the fuel is carried in the vehicle) and for any difference in static pressure of the jet exit of the engine (p_e) and the atmospheric or ambient pressure p_0 .

$$F = \dot{m}_a(v_e - v_0) + \dot{m}_p v_e + A_e(p_e - p_0)$$

The last term in this equation is called the pressure thrust; it is positive if $p_e > p_0$ (which occurs when there is incomplete expansion of the gases in the engine exit nozzle) or negative if $p_e < p_0$ (which occurs when there is overexpansion in the engine exit nozzle). For a rocket engine, the air mass flow $\dot{m}_a = 0$, and the thrust is equal to the last two terms of the equation only. In the case of a propeller engine, there are really two different air flows, and the first term in the above equation is split into two separate terms: one flow which crosses the plane of the propeller and is accelerated by the propeller blades and a second smaller airflow which goes through the engine to furnish oxygen for combustion. In the case of a nuclear

TABLE 1. ENERGY SOURCES AND WORKING FLUID FOR SEVERAL FLIGHT PROPULSION DEVICES

Type of Propellant	Type of Energy Source		
	Chemical	Nuclear	Solar
Surrounding medium used as working fluid	Aircraft piston engine with propeller	Nuclear turbojet Nuclear ramjet Nuclear submarine with propeller	--
Surrounding medium plus stored propellant or fuel	Turbojet Ramjet Blasted rocket	Nuclear turbojet with afterburning of chemical fuel	--
Propellant stored within vehicle	Liquid and/or solid propellant rocket Chemical battery with electric propulsion	Nuclear H_2 fission rocket Nuclear reactor power source with electrical propulsion	Solar heated hydrogen rocket
No propellant expelled	--	Future photon rocket	Solar sail

reactor energy source, $\dot{m}_p = 0$ and the second term can be omitted. In a vacuum $p_0 = 0$.

Consider a winged vehicle in equilibrium rectilinear flight in a two-dimensional (fixed plane) trajectory; assume all control forces, lateral forces and turning moments to be zero and the flight direction to be the same as the thrust direction. In the direction of the flight, the instantaneous vehicle mass m_v times the vehicle acceleration dv/dt has to equal the sum of all the forces, namely a component of the thrust F , the aerodynamic drag D , and a component of the gravitational attraction or the weight $m_v g$. The angles are as defined in Fig. 2.

$$m_v dv/dt = F - D - m_v g \sin \theta \quad (2)$$

The vehicle mass m_v multiplied by the acceleration in a direction perpendicular to the flight

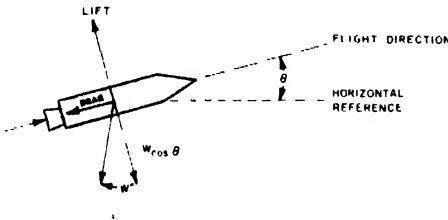


FIG. 2 Simple free body diagram of flying vehicle.

path ($v d\theta/dt$) must equal the sum of all forces perpendicular to the flight direction; here, the lift force L must be considered.

$$m_v v \frac{d\theta}{dt} = L - m_v g \cos \theta \quad (3)$$

The solution to these two equations results in the determination of a two-dimensional trajectory, maximum flight velocity, range, and other flight performance parameters. The actual solution is three-dimensional, and must usually be a numerical integration since m_v decreases with time, L and D vary with speed and altitude, and the direction of thrust is not the same as the flight direction; also, both the flight angle and the angle of attack are usually changing. For the case of a linear, simplified horizontal equilibrium flight, $\theta = 0$, $dv/dt = 0$, and thus Eq. (2) reduces to $F = D$. The vehicle mass m_v consists of the vehicle dry mass or final vehicle mass after expenditure of all propellant $(m_v)_f$ plus the propellant or fuel mass m_p . For steady fuel flow, $m_p = \dot{m}_p t$. For the case of gravity-free flight in a vacuum (true space environment), Eq. (1) can be rewritten for a rocket:

$$\frac{dv}{dt} = \frac{F}{m_v} = \frac{\dot{m}_p v_e}{(m_v)_f + m_p t}$$

Integration gives

$$v_v = v_e \ln \frac{m_v}{m_v - m_p} = v_e \ln \frac{(m_v)_f}{(m_v)_f - m_p}$$

Thus the maximum velocity attained by a rocket-

space vehicle operating in a gravitationless vacuum is equal to the product of the average effective rocket exhaust velocity v_e and a logarithmic function of the initial vehicle mass $(m_v)_f = (m_v)_f + m_p$ (fully fueled vehicle) at start of the engine operation, divided by the final vehicle mass (with all the fuel expended) $(m_v)_f$ at the end of engine operation. This velocity v_v will be large when v_e is large, i.e., high energy is available from the propellant or the engine and when $(m_v)_f$ is small, i.e., when the dry mass of the vehicle (dry engine mass, tanks, payload, or structure) is small and no unnecessary mass is designed into the vehicle. This means that m_p is large and the initial vehicle mass $(m_v)_f$ consists largely of propellant.

The *specific impulse* of an engine can be defined as the thrust force obtained from a unit propellant weight flow.

$$I_s = \frac{F}{\dot{w}} = \frac{F}{mg} \quad (\text{kg force/kg mass per second})$$

Since the propellant usually refers to the fuel stored in the vehicle and not to the air flow, the values of I_s for air-breathing engines are generally very high.

The *specific fuel consumption* (sfc) is usually expressed as the fuel flow rate (pounds per hour) per engine shaft brake horsepower. Both of these parameters are an indication of the quality of design and operation of an engine; a high value of I_s or a low value of (sfc) indicates efficient use of propellant or fuel.

Some of the most significant types of engines are described briefly in the remainder of this section.

Rocket Engines. These engines use both a fuel and oxidizing propellant and both are stored within the flying vehicle, making it independent from its surrounding fluid. Thus a rocket can operate in space, air, or under water. The supersonic nozzle jet exit velocity v_e of a rocket using ideal gas laws can be derived to be

$$v_e = \sqrt{\frac{2gkR}{(k-1)M} \frac{T}{p_c}} \left[1 - \left(\frac{p_e}{p_c} \right)^{(k-1)/k} \right]$$

where

v_e	nozzle exit velocity
g	gravitational constant
k	ratio of specific heats of gas
R	universal gas constant
M	molecular weight of hot gas
T	absolute combustion temperature
p_c	nozzle exit gas pressure
p_e	combustion chamber pressure.

The exhaust velocity (or the specific impulse which is $I_s = v_e/g$) increases as the molecular weight M is decreased or as the combustion temperature T is increased. Because of the pressure ratio effect, there is actually a slight (10 to 20 per cent) increase in specific impulse as the altitude is increased (lower ambient pressure) or as chamber pressure is increased. Air-breathing

engines, in comparison, lose specific impulse with altitude. Values of v_e or I_s calculated for a given propellant and engine from thermochemical and thermodynamic data are usually very close to actual performance (usually within 5 to 10 per cent), because rocket combustion efficiencies are high and nozzle losses are usually low. Schematic diagrams of several liquid and solid propellant systems are shown in Figs. 3 to 7, and some important applications are shown in Table 2.

In *liquid propellant rocket engines* the propellants are fed under pressure from tanks in the vehicle into a thrust chamber where they are injected, mixed, and burned at high pressures and very high temperatures to form the gaseous reaction products, which in turn, are accelerated in a nozzle and ejected at high velocities. The feed system for transferring the propellants into the thrust chamber includes valves and controls.

The principal components of a *thrust chamber* (Figs. 3 and 4) are the *nozzle*, the *chamber*, and the *injector*. An injector introduces and meters the flow of the liquid propellants and also atomizes and mixes them in the correct proportions in such a manner that they can be readily vaporized and burned. In the combustion chamber, the burning of the liquid propellant takes place at high pressure, usually between 5 and 150 atmospheres.

The *gas pressure feed system* (Fig. 3) offers one of the simplest and most common means of transferring propellants by displacing them with a high-pressure gas which is fed into the tanks under a regulated pressure. In a *turbopump feed system*, the propellant is pressurized by means of pumps driven by one or more turbines (Fig. 4) which derive their power from the expansion of hot gases. A separate gas generator ordinarily produces these gases in the required quantities

and at a temperature which will not hurt the turbine buckets (1200 to 2000°F).

Liquid propellants. A *bipropellant rocket unit* has two separate propellants, a fuel and an oxidizer

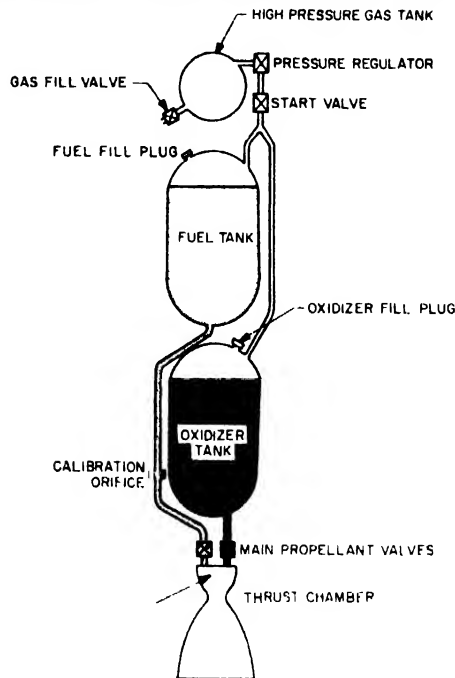


FIG. 3. Simplified schematic diagram of liquid propellant rocket engine with pressurized gas feed system and uncooled thrust chamber (reproduced by permission from McGraw-Hill Encyclopedia of Science and Technology, Vol. II, New York, McGraw-Hill Book Co.)

TABLE 2. TYPICAL DATA FOR VARIOUS ROCKET ENGINES

Type	Typical Range of Thrusts (lbs)	Typical Range of Duration	Application
High thrust liquid propellant rocket	1,000,000 to 4,000,000 for each engine with several engines in a cluster	1 to 5 min	Booster and sustainer stages of large missiles and space vehicles
Large solid propellant rocket	5,000 to 3,000,000	1 to 60 sec	Large and small missiles (surface to surface, surface to air, air to air, air to surface)
Prepackage storable liquid propellant	100 to 100,000	1 to 60 sec	Special applications of small missiles, lunar landing and takeoff
Jet assisted takeoff (solid propellant)	200 to 10,000	5 to 30 sec	Assist takeoff of airplanes
Space vehicle attitude control	1 to 150	0.01 to 10 sec/cycle; accumulate up to an hour	Control position, angle and orientation of spacecraft

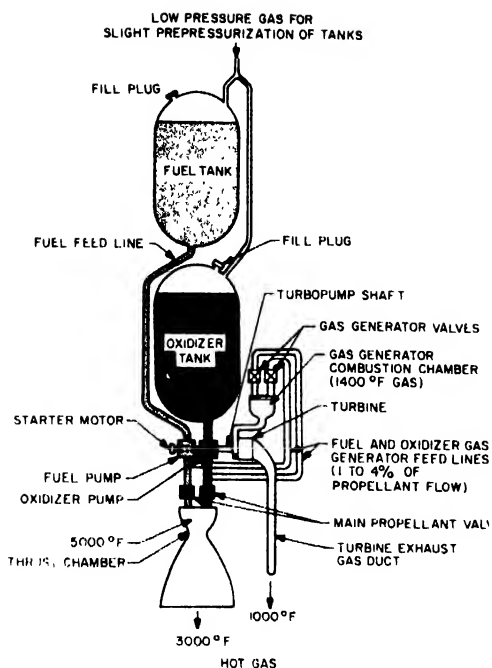


FIG. 4. Simplified schematic diagram of liquid propellant rocket engine with turbopump feed system and regeneratively cooled thrust chamber (reproduced by permission from McGraw-Hill Encyclopedia of Science and Technology, Vol. II, New York, McGraw-Hill Book Co.)

(such as kerosene and liquid oxygen), which are not mixed until they come in contact with each

other in the combustion chamber. Most liquid propellant rockets have been of this bipropellant type. A *monopropellant* contains oxidizing agent and combustible matter in a single substance. It can be a mixture of compounds, such as hydrogen peroxide with liquid alcohol, or it may be a homogeneous chemical agent, such as nitromethane. Certain types of liquid fuels and oxidizers are self-igniting and start burning when they come in contact with each other. Other types of propellants, for example, oxygen and alcohol, are not spontaneously ignitable, but require an igniter to furnish thermal energy for starting their combustion reaction. Typical values for liquid and solid propellants are given in Table 3.

Solid Propellants. In a *solid propellant rocket engine* all the propellant is contained within the combustion chamber. The hardware includes, in addition to the combustion chamber nozzle, an igniter and provisions for mounting the rocket (Figs. 6 and 7). Solid propellants themselves usually have a plastic, cakelike appearance (specific gravity is approximately 1.6) and burn at high pressure (10 to 150 atmospheres) on their exposed surfaces to form hot exhaust gases which are ejected through the nozzle. The physical mass or body of the propellant is called the *grain*. In some rockets, there is more than one grain inside the same combustion chamber.

The solid propellant grain contains all the material necessary for sustaining combustion. It can be a mixture of several chemicals, e.g., a mixture of ammonium perchlorate in a matrix of organic polymeric fuel such as rubber. Or it can be a homogeneous charge of special oxidizing organic chemicals, such as nitrocellulose or nitroglycerine. Once the propellant is ignited,

TABLE 3. TYPICAL THEORETICAL PERFORMANCE OF VARIOUS ROCKET PROPELLANTS

Propellant	Theoretical Thrust Chamber Specific Impulse at 500 psi Chamber Pressure, sec.		Bulk Specific Gravity	Optimum Mixture Ratio (Oxidizer-Fuel)	Combustion Temperature, F	Molecular Weight of Exhaust Gas lb/mole	Burning Rate in/sec
	At Sea Level, Area Ratio = 8	In Vacuum, Area Ratio = 25					
<u>Cryogenic Liquid</u>							
Oxygen and kerosene	261	324	63	2.25	5800	22	
Oxygen and 92.5% ethyl alcohol	249	311	61	1.5	5370	23	
<u>High Energy Liquid</u>							
Fluorine and Hydrogen	364	447	19	4.0	4700	9	
Fluorine and Hydrazine	303	372	80	1.75	7300	18	
Oxygen and Hydrogen	358	440	8	4.0	4500	8	
<u>Storable Liquid</u>							
Nitric acid and dimethyl hydrazine	246	304	76	2.4	5100	22	
90% hydrogen peroxide and kerosene			79	7.0	4600	22	
<u>Monopropellant (Liquid)</u>							
90% hydrogen peroxide and hydrazine	137	167	87	0	1365 1800	21	
<u>Solid</u>							
Double Base Solid	190-250	235 to 315	100	-	3000-5000	24	.2 to .9
Composite Perchlorate	180-250	220 to 315	100	-	3000-5000	24	.1 to .8
Nitrate Composite	175-210	210 to 255	100	-	2400-3000	21	.05 to .15

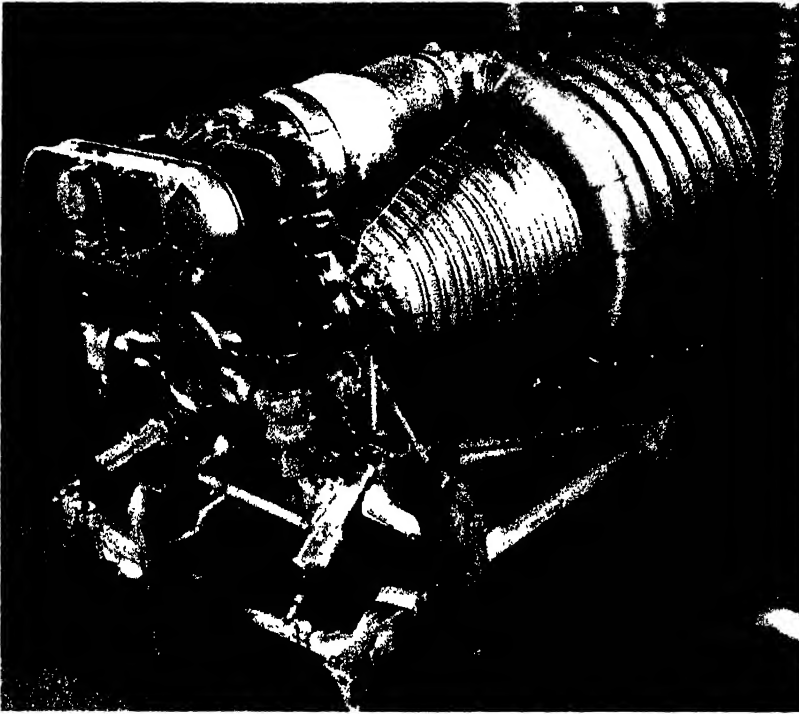


FIG. 5. F-1 rocket engine used in booster stage of advanced Saturn space vehicle.
(Courtesy of Rocketdyne, A Division of North American Aviation, Inc.)

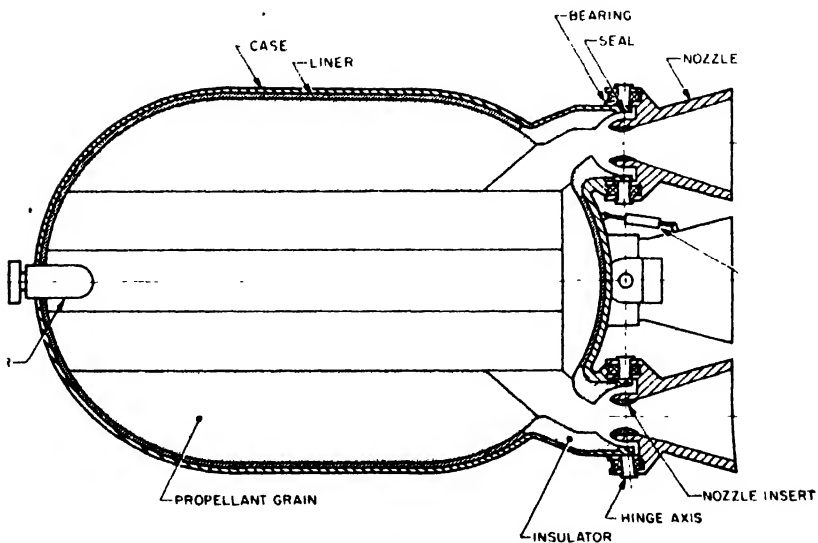


FIG. 6. Simplified schematic diagram of solid propellant rocket engine (reproduced by permission from G. P. Sutton, "Rocket Propulsion Elements," Third edition, New York, J. Wiley & Sons, Inc., 1963).

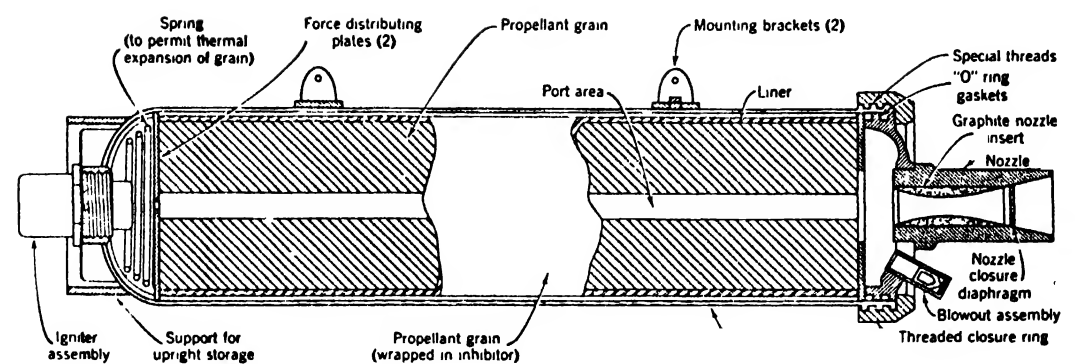


FIG. 7. Solid propellant rocket (reproduced by permission from G. P. Sutton, "Rocket Propulsion Elements" Third edition, New York, J. Wiley & Sons, Inc., 1963).

the grain burns smoothly on all of its exposed surfaces in a direction normal to the burning surface.

Compared to liquid propellant, the solid propellant requires no pump or pressurization for the fuel tank, and hence is mechanically simpler. The combustion chamber with solid propellant is larger, especially for large rockets, and is frequently operated at higher pressure than for a liquid engine. Solid rockets are simpler, more storable, and are usually more immediately ready for use, but are generally lower in performance when compared to liquid propellant units.

Most solid propellant grains are somewhat sensitive to variations in the ambient temperature and have a tendency to become soft on very hot days or brittle on very cold days; for certain propellants, therefore, it is necessary to restrict the temperature range over which they may operate. If the propellant should crack from too great a temperature variation (which induces thermal stresses) or from rough handling, additional burning surfaces would be created in the cracks and an unregulated increase in the pressure would cause failure of the chamber.

There is a maximum pressure above which smooth combustion can no longer be sustained and detonations may occur, and a minimum pressure below which stable, consistently smooth burning does not seem possible. Because of the active nature of the chemicals, some propellants deteriorate in storage. This can often be prevented by the addition of chemical stabilizers or inhibitors to the propellant.

In addition to producing a thrust force, solid and liquid propellant engines can be used also for producing auxiliary power and control torques to be applied to the vehicle. This latter is called *thrust vector control* and basically it is usually a mechanical means for altering the direction of the engine's thrust during flight.

Air-breathing Engines. As shown by Figs. 8 to 11, there exist a variety of different types of air-breathing propulsion engines. The range of performance values shown in Table 4 for different types of air-breathing engines are representative and do not correspond to data for specific engines. Each engine is optimized for a specific flight operating condition of speed (Mach number) and altitude. For example the turbojet

TABLE 4. TYPICAL DATA FOR AIRBREATHING ENGINES

Engine Type	Flight Speed (mph)	Altitude (feet)	Cruise Specific Fuel Consumption or Specific Thrust	Thrust to Weight or Power to Wt. Ratio	Typical Applications
Reciprocating Piston engine with propeller	0 - 400 mph	0 - 40,000	0.37 to 0.52 lb hp - hr	1.0 to 2.0	Transport aircraft Small airplanes Helicopters Target drones
Turbojet	0 - 2500	0 - 100,000	1.0 lb lb-hr at M 0.8 1.5 lb lb-hr at M 3.0	0.8 to 1.0 (100 to 400% increase with afterburner)	Bomber, Fighter, Transport aircraft, Target drones
Turboprop	0 - 550	0 - 50,000	0.8 lb hp-hr at M 0.6 0.25 lb hp-hr at M 0	2.0	Fast transport aircraft Small airplanes Helicopters
Ramjet	700 to 4200	0 - 120,000	1.5 lb lb-hr at M 2 2.0 lb lb-hr at M 6	up to 20.0	Target drones Anti-aircraft missiles

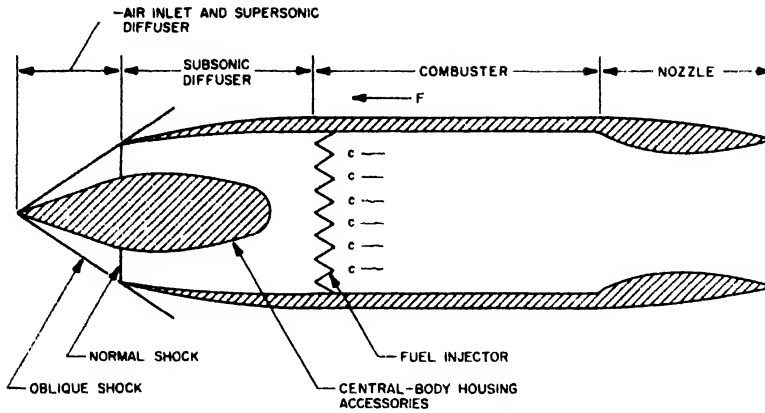


FIG. 8. Simplified schematic diagram of ram jet (reproduced by permission from H. H. Koelle, "Handbook of Astronautical Engineering," New York, McGraw-Hill Book Co., 1961).

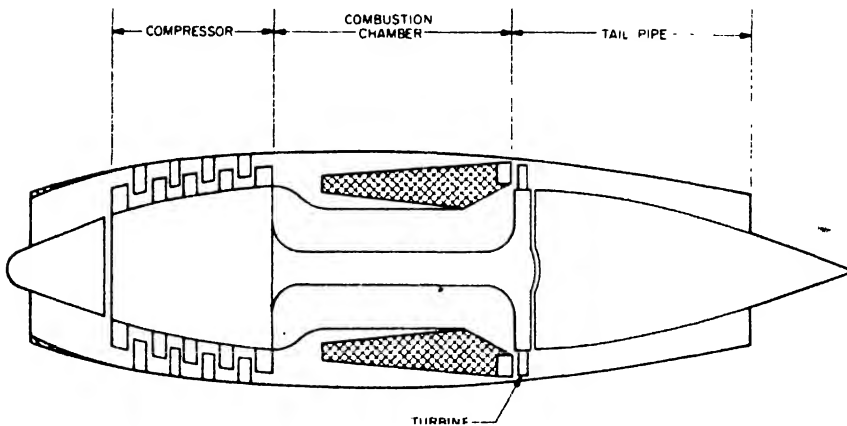


FIG. 9. Simplified schematic diagram of turbojet (reproduced by permission from H. H. Koelle, "Handbook of Astronautical Engineering," New York, McGraw-Hill Book Co., 1961).

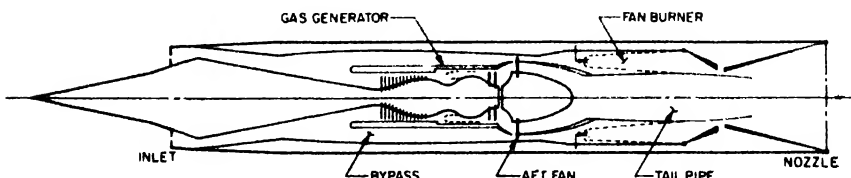


FIG. 10. Simplified Schematic diagram of advanced turbojet with bypass (reproduced by permission from H. H. Koelle, "Handbook of Astronautical Engineering," New York, McGraw-Hill Book Co., 1961).

reducing the velocity. After heating the flow in a combustion chamber, the reverse process occurs adiabatically in the nozzle, where it is desired to attain a maximum exhaust velocity. The efficiency of energy conversion in the inlet duct, nozzle, combustors, compressors, and turbines is a very important factor, and becomes a predominant criterion at supersonic and hypersonic velocities, when compression is achieved usually by a series of oblique shock waves commencing at an inlet spike. To maintain good efficiency and the desired airflow, some diffusers and nozzles incorporate a variable wall contour or cross-section geometry. In general, air-breathing engines have been well developed to a high state of reliability and have given millions of hours of good service.

The available oxygen from the air limits the combustion process. For example, at constant flight speed, the thrust thus decreases with altitude (or oxygen density) and below a combustion pressure of approximately 3 psi, combustion is not easily sustained (flameout limit). Available high-temperature materials will set an upper limit to the maximum combustion temperature at approximately 1700 to 2400 °F. At high speeds and high altitude, the ram-compression of the air causes its temperature to rise substantially, so that the amount of energy that can be added by combustion (without damaging turbine materials) is thus limited; also, special cooling provisions are required.

The simplest air-breathing engine is a *ramjet* (Fig. 8). It does not produce static thrust (at zero flight speed, such as during takeoff) and thus needs a rocket engine or some other engine to bring it to its minimum operating speed.

The *reciprocating piston engine with a propeller* was the very first engine to fly. It is the most economic engine for subsonic flight speed and is used in airplanes and helicopters. The hot gases do work against a piston (not a turbine), which in turn requires a crank mechanism to convert the reciprocating piston motion into shaft rotation. The use of a variable pitch propeller and superchargers for precompression of the inlet air further increase the economy.

The *turbojet* (Figs. 9 and 11) can be designed for a variety of speeds, altitudes and thrust ranges. It often includes special design features such as afterburners (increases the thrust) or bypass duct arrangements with and without a separate compressor called the fan (this improves the fuel economy over certain performance ranges of speed and altitude). This may also include a secondary set of combustion chambers which add heat to the bypass air. An advanced bypass turbojet is shown in Fig. 10.

The *fuel* used is usually a narrow-cut petroleum refinery hydrocarbon product, having the approximate formula of $\text{CH}_{1.95}$. For advanced air-breathing engines, the use of liquid hydrogen fuel offers considerable performance gains.

Advanced Engines. The use of nuclear fission as a source of thermal energy for heating a working fluid promises to give ramjets, turbojets (which

require no combustible fuel) and novel rocket engines with superior specific impulses. The development of such engines faces problems in materials, matching of reactor energy level with the flow level, shielding of radiation, and afterheat disposal, and is currently being investigated. Such a rocket engine using liquid hydrogen as a working fluid and a solid core fission reactor is shown in Fig. 11; it may become useful for manned planetary flight. The use of a nuclear ramjet, now being experimentally investigated, will permit the eventual development of aircraft with theoretically unlimited endurance.

Fusion reactions may someday allow the development of a second type of nuclear engine. They avoid undesirable by-products and could release their energy directly to the working fluid. However, techniques for controlled fusion are yet to be developed.

The use of electrical energy offers further potential increases in rocket engine performance. Electrical propulsion units are discussed separately in ELECTRIC PROPULSION.

Photon propulsion. Solar sail engines rely on the reflection of photons from the sun (radiation pressure at the distance of the earth from the sun is about $5 \cdot 10^{-7}$ kg/m²). Although this force is limited by being directed only "away" from the sun, solar "sailing" can provide low thrusts, and attitude control, turning a spaceship completely around in a few hours. No working fluid is carried in the vehicle.

For the vehicle to carry its own light source, a photon rocket engine of adequate thrust would necessitate energies and techniques far beyond present capabilities.

G. P. SUTTON

References

1. Koelle, H. H., "Handbook of Astronautical Engineering," New York, McGraw Hill Book Co., 1961.
2. Sutton, G. P., "Rocket Propulsion Elements," Third edition, New York, J. Wiley & Sons, 1963.
3. Morgan, H. E., "Turbojet Fundamentals," Second edition, New York, McGraw-Hill Book Co., 1958.

Cross-references: AERODYNAMICS; ASTRODYNAMICS; ASTRONAUTICS, PHYSICS OF; DYNAMICS; ELECTRIC PROPULSION; IMPULSE AND MOMENTUM; MECHANICS.

FLUID DYNAMICS

Fluid dynamics is study of the motion of matter in the gas, liquid, plastic or plasma state. When restricted to flow of incompressible (i.e., constant density) fluids, it is called *hydrodynamics*; when dealing with electrically conducting fluids with magnetic fields present, it is called *magneto-fluid dynamics*; when dealing with practical problems of air flow past airplane wings, through ventilating equipment, etc., it is called *aerodynamics*.

Basically two fundamental approaches are employed: (1) continuum or field dynamics and (2) kinetic theory and nonequilibrium statistical mechanics.

In continuum dynamics, fluid properties—namely, velocity u_i , density ρ , pressure p (more generally stress), temperature T , viscosity, conductivity, etc.—are assumed to be physically meaningful functions of three spatial variables x_1, x_2, x_3 and time t . Two of the four basic equations relating these continuum variables (no electrical current, linear relation between residual stress and rate of strain with viscosity coefficients μ and η' are:

Continuity

$$\frac{\partial(\rho u_i)}{\partial x_i} + \frac{\partial \rho}{\partial t} = 0$$

Motion

$$\begin{aligned} \rho \frac{\partial u_i}{\partial t} + \rho u_j \frac{\partial u_i}{\partial x_j} &= -\frac{\partial p}{\partial x_i} + \eta' \frac{\partial^2 u_i}{\partial x_k \partial x_k} \\ &+ \mu \frac{\partial}{\partial x_j} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \frac{\partial u_i}{\partial x_k} \frac{\partial \eta'}{\partial x_k} \\ &+ \frac{\partial u_i}{\partial x_j} \frac{\partial \mu}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \frac{\partial \mu}{\partial x_j} \end{aligned}$$

In the above equations, the index notation used for subscripts is that i, j, k may be 1, 2 or 3 representing components of vectors along axes 1, 2 or 3, respectively, and in any term where an index occurs twice, a sum over products is implied. Two additional equations expressing the heat transfer and viscous transfer of energy in forms involving statements of the first and second laws of thermodynamics and the equation of state of the material are required. The four equations, together with boundary conditions, constitute formally a complete set of equations determining p, T, ρ and u_i . For complete discussion and additional equations pertaining to the more general case, see reference 1.

These partial differential equations are nonlinear and have no general solutions even for the most restrictive boundary conditions. Solutions are carried out for very idealized flows. Examples of particular solutions for selected geometrical boundaries are given below.

In Fig. 1, the special flow called Couette flow is indicated schematically. The flow is between parallel plates, lower plate at $y = 0$ at rest, upper plate at y_B moving with constant speed u_B in the x direction. Stress throughout the fluid is constant, given by $P_{xy} = \mu du/dy = \mu(u_B/y_B)$. This is pure shear flow and experimentally is often considered to define and measure the viscosity coefficient μ assumed constant for the homogeneous fluid. The velocity profile appearing at the right in the figure shows by velocity arrows of different length at the various positions y how the velocity varies with position. Steady flow (no dependence of any quantity on time), constant pressure, constant density and laminar flow are additional assumptions for Couette Flow. The flow is realized experimentally by confining the fluid in the narrow annulus between rotating concentric cylinders of nearly equal radius; the cylinders rotate at different speeds.

In Fig. 2, the special flow is in a pipe of uniform cross section, pressure is assumed to be constant across each cross section but to vary linearly with distance x along axis of pipe so $\frac{dp}{dx} = \frac{(p_1 - p_2)}{L}$.

Pistons driving the flow are assumed to be infinitely far away, so that the flow velocity, parallel to pipe axis, has the same dependence upon y and z for all x . The velocity profile is parabolic in both the two-dimensional case (infinite parallel plates) and in the circular cross-section case. Mean flow velocity u_m and viscosity coefficient μ are assumed constant; the flow is assumed steady and laminar. For a circular cross-section pipe of radius a , at any distance r from the center, $u = 2u_m(1 - r^2/a^2)$, and the volume passing a cross section per second is $Q = \pi a^2 u_m = \pi a^4 (p_1 - p_2) / 8\mu L$. Since these formulas do not

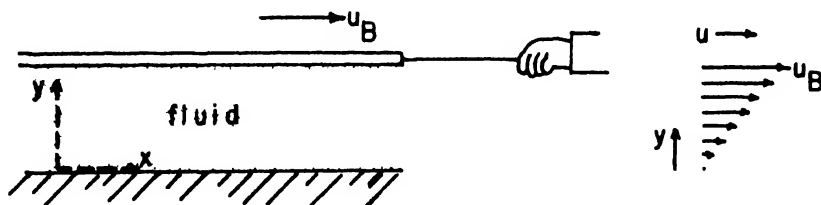


FIG. 1. Couette flow.

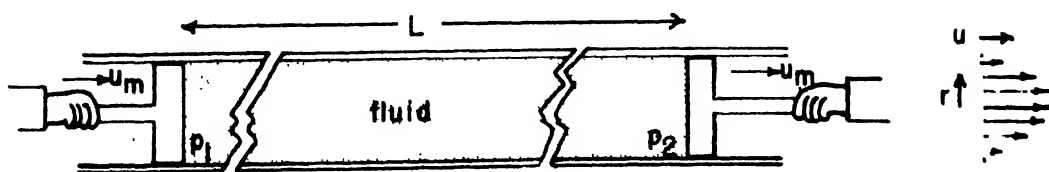


FIG. 2. Poiseuille flow.

apply near pipe entrances, caution in applying them to pipes of finite length is necessary even when the flow is steady and laminar. (See later discussion of turbulent and laminar flows.)

Other examples of idealized solutions are one-dimensional flow of an ideal gas through a normal shock wave, flow of an ideal gas without viscosity through a pipe of slowly changing cross section (wind tunnel), and one-dimensional finite waves in an ideal gas. Many other solutions involve making whatever approximations and assumptions are necessary to obtain descriptions of observed flows.

In kinetic theory and nonequilibrium statistical mechanics, fluid properties are associated with averages of properties of microscopic entities. Density, for example, is the average number of molecules per unit volume times the mass per molecule. While much of molecular theory in fluid dynamics aims to interpret processes already adequately described by the continuum approach, additional properties and processes are presented. The distribution of molecular velocities (i.e., how many molecules have each particular velocity), time-dependent adjustments of internal molecular motions, and momentum and energy transfer processes at boundaries are examples.

When motion of the fluid consists of only small fluctuations about a state of near-rest, the continuum equations are linearized by neglecting nonlinear terms and become the equations of acoustics². A large variety of fluid motions are described as sound waves; when the small-motion or acoustic description can be used, the *principle of superposition* is valid. This powerful principle allows addition of simple simultaneous motions to represent a more complex motion, such as the sound reaching the audience from the instruments of a symphony orchestra. The superposition principle does not apply to large-scale (nonacoustical) motions, and the subject fluid dynamics (in distinction from acoustics) treats nonlinear flows, i.e., those which cannot be described as superpositions of other flows. The description of small motions in a small region of even a nonlinear flow is useful; at each place in the flow, there is a "local sound speed."

Since sound waves travel with a speed relative to the fluid, waves moving in a moving fluid can sometimes be carried off in a direction opposite to the direction of sound travel. The flow where such a thing happens is called *supersonic*; the flow speed is greater than the sound speed at the spot where the flow is supersonic. Supersonic flow occurs around high-speed vehicles and missiles, and in pipes when high pressure gas escapes through a nozzle into a region of sufficiently lower pressure. A steady supersonic flow always must pass through a *shock front* to slow down to subsonic flow again.

The continuum description of flow fails to describe nearly all actual flows because actual flows when looked at carefully are *turbulent*. Turbulent flows have violent and erratic fluctuations of velocity and pressure which are not associated with any corresponding fluctuations of the

boundaries containing or driving the fluid. Turbulence is generally considered to be the manifestation of the nonlinear nature of the fundamental equations. Under certain conditions as mentioned earlier in describing Couette and Poiseuille flows, nonturbulent or *laminar* flow exists. A common example is cigarette smoke rising from a cigarette held at rest; near the cigarette, the stream is smooth and straight, or laminar, and further up the flow breaks into turbulence.

Reynolds showed that Poiseuille flow in a pipe occurs when $\rho u a / \mu$ is smaller than 2000. The combination of variables is dimensionless and is called the *pipe Reynolds number*. Blood flow in capillaries is laminar, but water flow in household pipes is turbulent unless the flow is about that allowed by a leaky faucet or less. Other types of flow have Reynolds numbers characterizing transition from laminar to turbulent; for example a sphere falling in a fluid of viscosity μ obeys *Stokes Law*

$$mg = 6\pi\mu u$$

where u is the constant speed of fall, m is the sphere mass, a its radius, and g the weight per unit mass, but the law is obeyed only if the *sphere Reynolds number* $\rho u a / \mu$ is smaller than about 1.

Because of turbulence and viscosity, the very simple and useful *Bernoulli formula* is not valid; it can be derived as applying to a constant-density fluid with zero viscosity in laminar flow. However, under certain conditions, the formula applies approximately even when the flow is turbulent, predicting properties within 5 to 20 per cent of the observed values. The Bernoulli formula states

$$p + \rho gh + \frac{1}{2} \rho u^2 \quad \text{same constant in all places in the fluid}$$

where p is the fluid pressure, ρ is the fluid mass density (which must be treated as constant), u is the fluid speed, h is the vertical height above some convenient reference level, and g is the weight per unit mass.

When combined with the equation of continuity, the Bernoulli formula gives a simple description of the Venturi, used in the automobile carburetor (see Fig. 3). Continuity states $u_1 A_1 = u_2 A_2$ volume crossing any cross section of the pipe per second. $u_1 / u_2 = A_2 / A_1$ is small so u_1 is smaller than u_2 . The Bernoulli formula states $p_1 + \frac{1}{2} \rho u_1^2 = p_2 + \frac{1}{2} \rho u_2^2$, and since u_1 is smaller than u_2 ,

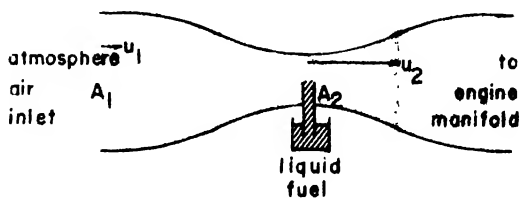


FIG. 3. Pipe flow with a constriction; carburetor employing Venturi.

p_1 is larger than p_2 . The atmospheric pressure p_1 pushes the liquid up into the lower pressure region p_2 .

Another common situation described by the Bernoulli formula is the discharge of (constant density) fluid from a small hole. For the cylindrical bucket of water in Fig. 4, equate the sum of the quantities in the Bernoulli formula at the top surface to the sum at the hole: $p_1 + \rho gh + 0 + p_2 + 0 + \frac{1}{2}\rho u_2^2$. The pressure at both top and bottom is atmospheric pressure p_1 ; the speed at the top is approximately zero because the hole is considered to be very small. The predicted speed of the emerging water is therefore $\sqrt{2gh}$ regardless of the size of the hole (as long as the hole is small).

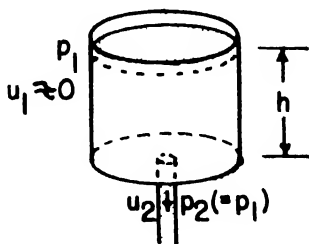


FIG. 4 Discharge through a small hole in a bucket

A valuable instrument in the form of a probe for observing fluid speed is the Pitot tube; its operation is described by the Bernoulli formula. A glass or metal tube with an open end points into the flow, and the pressure difference Δp between the stagnant fluid in the tube and the moving fluid allows calculation of the fluid speed at the place where the tube tip is inserted by $\sqrt{2\Delta p/\rho}$ where ρ is the fluid mass density. When observing air speed, the pressure difference Δp is easily measured by connecting the open ended tube via rubber hose to a glass U-tube water manometer. Errors as much as 50 per cent may easily occur in various practical situations, but order-of-magnitude measurements at least are usually possible.

RAYMOND J. EMRICH

References

1. Flügge, S., and Truesdell, C., Eds., "Encyclopedia of Physics," Vol. VIII, Parts 1 and 2, "Fluid Dynamics," Berlin, Springer-Verlag, 1959, 1963.
2. Hunt, F., "Propagation of Sound in Fluids," "American Institute of Physics Handbook," Article 3c, New York, McGraw-Hill Book Co., 1963.
3. Lamb, H., "Hydrodynamics," First American Edition, New York, Dover Publications, 1945.
4. Hirschfelder, J., Curtiss, C. F., and Bird, R. B., "Molecular Theory of Gases and Liquids," New York, John Wiley & Sons, 1954.
5. Bradley, J. N., "Shock Waves in Chemistry and Physics," New York, John Wiley & Sons, 1962.

Cross-references: FLUID STATICS, KINETIC THEORY, MAGNETO-FLUID-MECHANICS, STATISTICAL MECHANICS.

FLUID STATICS

Statics involves a study of the conditions under which a body remains at rest. If a body of fluid is at rest, the forces are in equilibrium or the fluid is in static equilibrium. The types of force which may act on a body are shear or tangential force, tensile force, and compressive force. Fluids move continuously under the action of shear or tangential forces. Thus, a fluid at rest is free in each part from shear forces; one fluid layer does not slide relative to an adjacent layer. An example is a quantity of water or liquid at rest in a bottle.

Fluids can be subjected to a compressive stress which is commonly called "pressure." The term pressure will be used to designate a force per unit area; pressure units may be dynes per square centimeter or pounds per square foot.

Atmospheric pressure is the force acting on a unit area due to the weight of the atmosphere. "Gage" pressure is the difference between the pressure of the fluid measured (at some point) and atmospheric pressure. The term vacuum refers to pressures below atmospheric. "Absolute" pressure, which can be measured by a mercury barometer, is the sum of gage pressure plus atmospheric pressure.

Pascal's law states that the pressure in a static fluid is the same in all directions. This condition is different from that for a stressed solid (as steel) in static equilibrium; in such a solid, the stress on a plane depends on the orientation of that plane. A liquid in contact with the atmosphere is sometimes called a "free surface." A static liquid has a horizontal free surface if gravity is the only type of force acting.

Imagine a body of static fluid in a gravitational field. The mass of the fluid is m (as in grams) and the weight of the fluid is mg (as dynes) where g is the local gravitational acceleration. Figure 1 indicates a large region of any static fluid with a very small or infinitesimal element. Figure 2

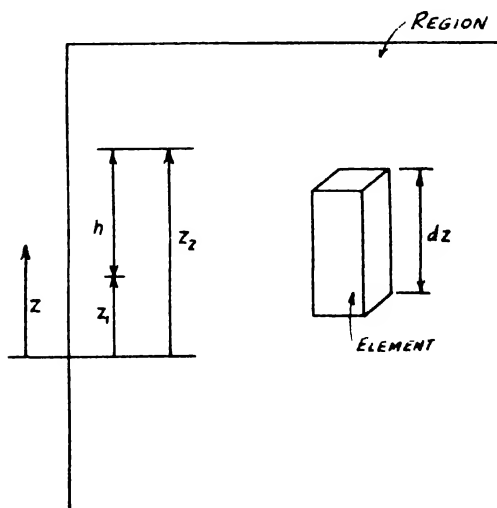


FIG. 1. Large region of any static fluid.

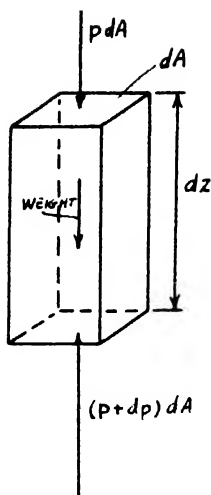


FIG. 2. Vertical forces on infinitesimal element.

indicates the element in detail. The vertical distance z is measured positively in the direction of decreasing pressure (up), dA is an infinitesimal area, p is the pressure acting on the top surface, and $(p + dp)$ is the pressure acting on the bottom surface; the pressure difference is due only to the weight of the fluid element. Let ρ represent density, which is mass per unit volume (as grams per cubic centimeter). Thus the weight of the element is $\rho g dz dA$. Considering the element as a free body, an accounting of forces in the vertical direction gives

$$dp dA = \rho g dz dA \quad dp = \rho g dz \quad (1)$$

As z is measured positively upward, the minus sign indicates that the pressure decreases with an increase in height. This fundamental equation of fluid statics can be applied to all fluids. In integral form, Eq. (1) becomes

$$\int_1^2 \frac{dp}{\rho g} = \int_1^2 dz \quad (z_2 - z_1) \quad (2)$$

where 1 refers to one level and 2 refers to another level. The functional relation between pressure p and the combination ρg must be established before Eq. (2) can be integrated. There are two major cases: (a) incompressible fluids, in which the density ρ is constant and (b) compressible fluids, in which the density ρ varies.

Liquids can be considered as incompressible in many cases. For small differences in height, a gas might be regarded as incompressible. For an incompressible fluid, with constant g , Eq. (2) becomes

$$p_2 - p_1 = \rho g(z_2 - z_1) \quad (3)$$

The term $(z_2 - z_1)$ may be called a static "pressure head," and it can be expressed in feet or inches of water, or some height of any liquid. For example, barometric pressure can be expressed in inches of mercury.

A "manometer" is a device that measures a static pressure by balancing the pressure with a column of liquid in static equilibrium. A large variety of manometers are used, such as differential, vertical, and inclined. The common mercury barometer is essentially a manometer for measuring atmospheric pressure; a mercury column in a glass tube balances the weight of the air above the mercury. Figure 3 illustrates a manometer in which the left leg is open to the atmosphere; the liquid has a specific weight $\rho_2 g$. In the other leg is a liquid of specific weight $\rho_1 g$. Starting with the left leg, the gage pressure p_A at point A is

$$p_A = h_2 \rho_2 g$$

Since the fluid is in static equilibrium, the pressure p_B at point B equals the pressure at point A. Thus

$$p_A = p_B = h_2 \rho_2 g$$

The pressure p_C at point C is less than that at B. Thus

$$p_B - p_C = h_1 \rho_1 g$$

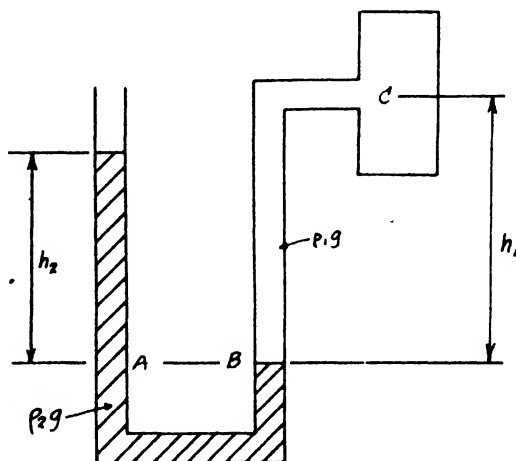


FIG. 3. Manometer.

Then the gage pressure at point C is

$$p_C = g(h_2 \rho_2 - h_1 \rho_1)$$

When a body of any kind is partly or completely immersed in a static fluid, every part of the body surface in contact with the fluid is pressed on by the fluid; the pressure is greater on the areas more deeply immersed. The resultant of all these fluid pressure forces is an upward or buoyant force. The pressure on each part of the body is independent of the body material. Archimedes principle states that the buoyant force equals the weight of the displaced fluid.

Equation (3) is for the special case of an incompressible fluid. As an example of a compressible fluid, consider an isothermal or constant-temperature layer of gas. The equation of state

for such a gas can be written

$$p = \rho RT_1 \quad (4)$$

where T_1 is the given absolute temperature and R is a gas constant or gas factor depending on the gas. Assuming a constant g , Eq. (2) gives

$$\frac{RT_1}{g} \int_1^2 \frac{dp}{p} = (z_2 - z_1) \quad (5)$$

$$z_2 - z_1 = \frac{RT_1}{g} \log_e \frac{p_1}{p_2}$$

Equation (5) is sometimes called a "barometric-height" relation. For an isothermal atmosphere a measurement of the temperature T_1 and the static pressure (as with a barometer) at two different levels will provide data for the calculation of the height difference.

RAYMOND C. BINDER

References

- Binder, R. C., "Fluid Mechanics," Englewood Cliffs, N. J., Prentice-Hall, 1962.
Prandtl, L., and Tietjens, O. G., "Fundamentals of Hydro- and Aeromechanics," New York, McGraw-Hill Book Co., 1934.

Cross-references: FLUID DYNAMICS, MECHANICS, STATICS

FORCE. *See* STATICS.

FRICTION

Introduction. Friction is the resistance to motion which exists when a solid object is moved tangentially with respect to the surface of another which it touches, or when an attempt is made to produce such motion. Friction thus takes its place as one of the general systems of force which are considered in MECHANICS (others being GRAVITY, ELASTICITY, etc.). Unfortunately friction depends to a marked extent on the material properties of the contacting surfaces, and even more importantly, on any surface contaminants which may be present, so that it is very difficult to estimate values of the friction force theoretically, with an uncertainty of less than about 20 per cent. In many calculations in solid mechanics, this uncertainty in the friction constitutes the limiting factor in determining the accuracy of the over-all calculation.

The friction force arises from the fact that, when two solids are pressed together, bonding between their surface atoms occurs, and these bonds have to be broken before sliding can commence. Bonding of any considerable strength occurs only in places where the surface atoms come within range of each other's strong force-fields (i.e., closer than about 3×10^{-8} cm) thus, when ordinary solids with appreciable roughness are used, bonding is confined to a few small patches (called junctions) over their inter-

face, where the high spots (or asperities) of one material have made contact with asperities on the other material.

The energy used up in the friction process appears almost entirely in the form of heat. Generally this consists of a moderate temperature rise over the contacting bodies, and superposed on this there are higher "flash temperatures" at the junctions. At high sliding speeds, softening or even melting of the tips of the asperities may occur.

In addition to the bonding or "adhesion" effect, which is the principal cause of friction, there are four other mechanisms which use up energy during sliding; energy which must be supplied by the friction force. These mechanisms are:

- (1) A roughness effect, caused by the interlocking of asperities and the need to lift one surface over the high spots of the other;
- (2) A ploughing effect, whereby an asperity on a hard material can dig a groove in a softer material;
- (3) A hysteresis effect, whereby there is elastic and plastic deformation of the material at or near the junctions, and not all the deformation energy is recoverable;
- (4) An electrostatic effect (with electric insulators), where work must be done to separate electrically charged regions on the sliding surfaces.

In a great majority of applications these four mechanisms do not account for as much as 20 per cent of the total resistance to sliding. The widely held belief that friction is due mainly to a roughness effect does not find experimental support. Cleaved mica (smooth to an atomic scale) shows very high friction.

Laws of Friction. If a normal load L presses two surfaces together, then we may apply a tangential force up to some limiting value F_s , and the surfaces will remain at rest. Sliding occurs when the tangential force exceeds the static friction force F_s , and almost as soon as motion starts the tangential force takes on a characteristic value, F_k , and always acts in a direction opposite to the relative velocity of the surfaces. F_s is often some 30 per cent larger than F_k , but sometimes they are equal.

The ratio F_s/L is the static friction coefficient f_s (or μ_s), while the ratio F_k/L is the kinetic friction coefficient f_k . We may cite quite general statements, or "laws," involving these coefficients of friction. The three classical laws, dating back to the seventeenth and eighteenth centuries, are:

- (1) The friction coefficient is independent of the load.
- (2) The friction coefficient is independent of the contact area.
- (3) The kinetic friction coefficient is independent of the sliding velocity.

More recently, it has been found possible to make another general statement about the friction coefficient:

- (4) The friction coefficient is independent of the surface roughness.

We may summarise these laws in the comprehensive statement.

(5) The friction coefficient is essentially a material property of the contacting surfaces, and of the contaminants and other films at their interface.

Although in practice these "laws" are reasonably well obeyed, there are often systematic divergencies, some of which have important consequences. In the rare cases in which the friction coefficient varies with load, it is often because of some special effect (e.g., a surface coating which is broken up at heavy loads); however, materials with a high elastic limit (e.g., polymers) generally show a friction coefficient which goes down somewhat as the load is increased. The fourth law applies closely to the intermediate ranges of surface roughness, but it is found that very smooth surfaces have higher friction because they tend to seize, and very rough surfaces have higher friction because of asperity interlocking.

Velocity and Time Effects in Sliding. Violations of the third friction law are important because they can lead to friction-induced oscillations. With some materials f_k is almost independent of velocity over a very wide range; however, with hard materials the friction generally goes down as the speed goes up (a so-called negative characteristic), while soft materials show a negative characteristic at high sliding speeds and a positive characteristic at low sliding speeds. Often, f_k changes by about 10 per cent during a factor-of-ten change of velocity. In a sliding system which has elastic compliance, a negative characteristic introduces an instability, and at high speeds this takes the form of harmonic oscillations (e.g., a violin string), while at low speeds relaxation oscillations occur (e.g., a creaking door). The relaxation oscillations are usually referred to as stick-slip, because during a part of each cycle the surfaces are at rest. The severity of the stick-slip is enhanced because, while the surfaces are at rest, f_s increases somewhat with time of stick (about 10 per cent for every factor-of-ten increase in time of stick above 10^{-1} seconds).

Value of the Friction Coefficient. In most situations the total real area of contact, or sum of all the asperities, is produced by plastic deformation of the asperities, and if we assume that strong bonds are formed joining the materials across the interface we find that the friction

coefficient is given by

$$f = s/p$$

where s is the plastic shear stress, and p the plastic indentation hardness, of the softer of the two contacting materials. Since s and p are similar plastic strength parameters, their ratio tends to be constant, within the range 0.3 to 0.6, for a wide variety of materials. With some materials the real area of contact increases during sliding, and friction coefficients above 1.0 are then commonly observed. Especially bad in this connection are materials with a high ratio of surface energy to hardness, namely clean soft metals such as lead, aluminum and copper.

A lubricant acts by introducing a layer of lower shear stress s_1 at the interface, and lowers the friction coefficient accordingly, down to 0.05 in favorable cases. Many nominally clean materials also give low friction, as result of contamination during manufacture or handling. Slight differences in the degree of contamination can produce drastic differences in the friction.

Some typical friction values are given in Table I.

Materials with Unusual Frictional Properties. In this category we may place the hard non-metals (diamond), which give low friction (~ 0.1); the elastomers (natural rubber) which give very high friction (~ 0.9); the layer-lattice substances (graphite, molybdenum disulfide, cadmium iodide) which give low friction (~ 0.1) and are used as solid film lubricants; the hexagonal close-packed metals (zinc, magnesium, titanium, cobalt) which when clean give lower friction coefficients (~ 0.5) than do other metals; "Teflon," which adheres very poorly to other solids and accordingly gives very low friction (~ 0.05); and metals in group IVb of the periodic table (titanium, zirconium, hafnium) which cannot be well lubricated by any known liquid substances. Generally, the frictional anomalies of solids may be traced to peculiarities in their structure or surface properties.

Related Fields of Interest. Friction is considered one of a group of mechanical surface interaction phenomena, and research in the other members of the group is of great interest to workers in the friction field. There is an extensive common literature. The other interaction phenomena are wear (the removal of surface material as result of mechanical action), lubrication (the properties

TABLE I. COEFFICIENTS OF FRICTION

Materials	Surface Conditions	f_s	f_k
Metals on metals (e.g. steel on steel, copper on aluminum)	Carefully cleaned	0.4-2.0	0.3-1.0
	Unlubricated, but not cleaned	0.2-0.4	0.15-0.3
	Well lubricated	0.05-0.12	0.05-0.12
Nonmetals on nonmetals (e.g., leather on wood, rubber on concrete)	Unlubricated	0.4-0.9	0.3-0.8
	Well lubricated	0.1-0.2	0.1-0.15
Metals on nonmetals	Unlubricated	0.4-0.6	0.3-0.5
	Well lubricated	0.05-0.12	0.05-0.12

of surface films which reduce friction and wear), and adhesion (the tendency of solid objects pressed together to remain together).

ERNEST RABINOWICZ

References

- Bowden, F. P., and Tabor, D., "Friction and Lubrication of Solids," Oxford, Clarendon Press, Part I, 1950, Part II, 1964.
Rabinowicz, E., "Friction and Wear of Materials," New York, Wiley, 1965.

FUSION

Introduction. When nuclei of certain light elements fuse together, there is an excess binding energy which appears in the form of heat. This is the energy source of both the sun and the hydrogen bomb, and it is the object of fusion research to liberate the energy in a controlled manner on earth; because of the abundance of light elements, such power is virtually unlimited.

In order to promote nuclear fusion, the reacting nuclei must have sufficient inertia to overcome their mutual coulomb repulsion. Thus nuclei with a high ratio of mass to charge, such as deuterium and tritium, heated to a high temperature are most suitable. The reaction of interest is



where the numbers in parenthesis are the energies of the reaction products. The cross section (σ) peaks at about $5 \times 10^{-24} \text{ cm}^2$ and deuteron energy of 107 KeV, liberating a total energy of 17.6 MeV and giving an energy gain of 160. This is emphasized by noting that 8 gallons of seawater contain 1 gram of deuterium, which yields a maximum of $8 \cdot 10^{10}$ calories on fusion, equivalent to 2500 gallons of gasoline or 80 tons of TNT. Many methods have been proposed, but are as yet untested, for harnessing this energy. Among these are the generation of electric current by the interaction of the charged helium particles with a magnetic field, the dissipation of the neutron energy as heat in a water jacket, using the neutrons to release fission energy in uranium placed round the fusion reactor (making a combined fusion, fission reactor), and the generation of steam power by using the heat radiated from the hot gases.

Although the cross section peaks at a directed deuteron energy of 107 KeV, it is possible to use a much lower mean energy if the nuclei are thermalized. This is because there is an exceedingly sharp rise of cross section with energy and most of the fusion reactions are produced by nuclei with energies far in the tail of the Maxwellian distribution. For example, in a thermal mixture of deuterium and tritium most of the reactions are produced by deuterium of energy $1.3 \times 10^{-4} T^{2/3} \text{ KeV}$ (where $T^\circ \text{ K}$ is the temperature). Even so, the temperature required is extremely high, typically near 10^8° K .

Plasma in a Fusion Reactor. At such high temperature, atoms are ionized by collisions and the electrons are not bound to any particular nucleus; this state of matter is called a plasma. Since stars are in the plasma state, plasmas are by far the most common form of matter in the universe (though they are rare on earth) and the problems of controlled thermonuclear fusion are closer to astrophysical plasma physics than nuclear physics.

Because of the unbound electrons, plasmas are good electrical conductors and a magnetic field cannot easily diffuse through them (the "skin depth" in a hydrogen plasma after a time t is $3 \times 10^6 t^{1/2} T^{-1/2} \text{ cm}$). While the diffusion proceeds there is a current induced in the plasma skin which interacts with the field to cause a pressure perpendicular to the field lines of $\sim B^2/8\pi$ atmospheres, where B is the field in kilogauss; effectively the plasma pressure is balanced by the field pressure which is supported by the magnet coils generating the field. Thus the magnetic field affords the means of insulating the plasma from the material walls against which it would rapidly cool.

Often, as the result of the method of forming the plasma or because of field diffusion, magnetic field becomes mixed with the plasma. The confining magnetic field pressure must then balance the sum of the plasma pressure and the pressure of the field in the plasma. In this connection, an important parameter of the plasma is β , the fraction of the confining field pressure that supports the plasma pressure, which varies from 0 to 1 as the field inside the plasma goes from the confining field to zero. For a particular plasma pressure, the lower the β , the more the confining field and the more the energy required to operate the magnet.

Operating Regime. The operating regime of a fusion reactor is determined by the variation of the reaction cross section with temperature, the economics of energy conversion into useful work, and material properties such as the strength of the magnet metal, vaporization of the wall of the containment vessel, etc. Energy balance in the reactor gives a minimum requirement for the product nt where t is the reaction time and n is the number of nuclei per cubic centimeter. If the total energy released in the reaction vessel (the nuclear energy plus the energy of the plasma particles plus energy radiated) is returned with an efficiency of 1/3 to maintain the plasma temperature and make up the radiation losses, then nt must be greater than 10^{14} and the minimum temperature is about $3 \times 10^7^\circ \text{ K}$. For this minimum condition, 1 per cent of the nuclei fuse together. Also, since the reaction rate depends on n^2 , too high an n rapidly leads to wall vaporization and too low an n to a negligible rate. If we assume that a power density of 100 watts/cc can be handled continuously (equal to present fission reactor levels), n must be $\approx 10^{16}$ nuclei/cc (about a thousandth of an atmosphere); this is probably correct to within an order of magnitude, being on the low side for pulse operated, highly

compressed plasmas that fill only a small fraction of the reaction vessel. The condition $nt > 10^{14}$ then demands that $t > 10$ msec, and for a $\beta = 1$ plasma a confining field of 34 kilogauss.

If the helium from the fusion reaction is trapped in the plasma, its energy can maintain the plasma temperature against losses by radiation. This avoids the inefficiency of recycling the fusion energy to heat more plasma. The minimum temperature for this approach, called the "ideal ignition temperature," is 4.6×10^7 °K. All the energy of the neutrons from the fusion reaction can then be applied to produce useful power.

Radiation. One of the many problems to be solved in devising a fusion reactor is that of radiation at these high temperatures. Fortunately, the plasma required is so diffuse that it does not act as a blackbody which would follow a fourth-power law of temperature for the radiant energy ($6 \times 10^{-13} T^4$ watts/cm²). Instead, the radiation mean free path is larger than the plasma dimensions and the dominant radiation process is caused by electrons deflected in the coulomb field of the nuclei. This is called bremsstrahlung and peaks in the soft x-ray region at a wavelength of around 1 Å. The power radiated is $0.54 \times 10^{-30} Z^2 n_e n_i T^{1/2}$ watts/cc (Z = nuclear charge; n_e and n_i are electron and nuclear density) which is proportional to the square root of the temperature and, for reasonable sized reactors, is far less than blackbody radiation. However, even hydrogenic bremsstrahlung is considerable and an increase of two orders of magnitude could kill the concept of fusion reactors. Because of the $n_e Z^2$ factor, ionized impurities could produce a serious radiation loss; for instance, only 1 per cent of ionized oxygen increases the radiation by 77 per cent. Impurities also absorb a disastrous quantity of energy during ionization, not only because of the ionization potential of the electrons but also in the excitation of the partially ionized atoms by free electrons with insufficient energy to produce complete ionization; this excitation energy is radiated a few nanoseconds after excitation. It is therefore very important to ionize impurities rapidly.

A further important mechanism of energy loss in the plasma is charge exchange produced between a neutral impurity molecule (liberated, for instance, at the wall of the vacuum vessel by the radiation) drifting into the plasma and donating an electron to a hot ion. The hot ion now becomes neutral, is no longer held by the magnetic field, and is lost from the plasma. Meanwhile the cold impurity ion absorbs plasma energy in being further ionized and heated, and the hot neutral particle bombards the wall to liberate further impurities.

Therefore because of both radiation and charge exchange every effort is made to produce a highly pure plasma, typically with an impurity level of less than 0.1 per cent. This is difficult considering the low operating density and the high temperatures involved.

Plasma Heating. For temperatures up to a few million degrees, the conductivity of the plasma

is sufficiently small for ohmic heating to be effective. Above these temperatures, the conductivity ($8 \times 10^{-6} T^{3/2}$ mhos/cm) becomes too large and other forms of heating must be used. Ohmic heating is often used as preheating so that when a magnetic field is later applied to the plasma surface it exerts a pressure rather than diffusing into the plasma.

A common method of heating is to compress the plasma with the magnetic field; the energy put into the plasma is then

$$\int_{V_1}^{V_2} p \, dv$$

where p is the magnetic field pressure and $V_1 - V_2$ is the volume change. The volume change is limited by the size of the magnet, so it is desirable to use a high value of p . In the "Theta" and "Z" pinch experiments the gas is heated in a cylindrical tube first ohmically, then by a powerful magnetic pressure. The magnetic field is produced by current from a very low inductance, high voltage capacitor energy store. The compression is then so rapid that shock fronts develop in the plasma; subsequently isentropic compression is produced by raising the field still further. Temperatures of the order of 10^7 °K are commonly quoted, and even 10^8 °K has been claimed for late in the life of the plasma. A method of obtaining high plasma compressions without large volumes of magnetic fields is to transfer the plasma into successively smaller magnets as the compression increases. The transfer is achieved in an experiment called "Toy Top" by using a larger field at one end of the plasma than at the other.

Another method of heating the plasma by magnetic compression without large volume changes is magnetic pumping. The plasma is alternately compressed and decompressed at one point on its length, and the problem is to choose conditions in which the plasma does not cool on decompression (that is, an irreversible cycle is required). The heating depends on the pumping period compared to the ion/ion collision time and the transit time of an ion through the pumping region. If the pumping period is of the order of the transit time and much shorter than the collision time, the ions gain energy while the field is increasing and leave the region before decompression. The heating is then proportional to $T^{3/2}$ and becomes more effective at high temperatures. This system and ion cyclotron heating (another system depending on oscillating magnetic fields) are being tested on the "Stellarator" experiment to bring ohmically heated plasma up to thermonuclear temperatures.

A third important means of forming the plasma is to accelerate individual nuclei in an electrostatic field, then inject them into the confining magnetic field. This has the advantage that high particle energies (on the order of tens of keV) are relatively easy to produce, but the problem of injecting the nuclei is difficult because it is not possible in general to trap the nuclei in a static magnetic field under conservative conditions. The

procedure is to change the particle orbit discontinuously either by a collision against background plasma or by altering the charge-to-mass ratio. For collision with background plasma, a very long path length is required: in the "Sinelnikov" experiment this is achieved by carefully angling the injection down a long cylindrical field with regular ripples along its length and high "mirror" magnetic fields at the ends, so that the particles are reflected back and forth between the mirrors and the ripples cause the particles on their first return to avoid the injection gun. In the "D.C.X. Mirror" experiment, fast D_2^+ molecules are injected, and dissociated in an arc. In the "Phoenix Mirror" and "Alice Mirror" experiments, excited neutral atoms are injected which are ripped apart when they reach the magnetic field because, due to the difference in charge, the negative electrons and positive ions in the atom gyrate in opposite directions (this is called "Lorentz" trapping).

Confinement. One of the difficulties of confining a plasma with a magnetic field is to eliminate end loss down the plasma axis. One way is to increase the magnetic field at the plasma ends so that the field lines bend towards the axis. In a collision-dominated high-density plasma, a mode of operation of the "Ft-1" pinch, the system then acts like a gas trapped in a double ended rubber balloon. Due to the curvature at the ends there is a component of force parallel to the axis which reflects the particles; this is called a collision-dominated magnetic mirror. However, even a very small amount of field accidentally mixed in the plasma will prevent the ends from closing completely and allow particles to escape. It is estimated that a field in the plasma of only 1 per cent (equivalent to $\beta = 0.999$) will produce an unacceptable loss.

Another type of magnetic mirror is the collision-free mirror. This is used with a low- β plasma when the ions gyrate around the magnetic field lines and move along them because of axial components of velocity gained by collisions. When the particles reach the collision-free mirror region there is again a component of magnetic force parallel to the axis which, if the axial energy is not too great, will reflect the particle back to the main plasma volume. A certain fraction, depending on the coulomb collision cross section, of the particles gains axial energy sufficient to overcome the mirror and is lost. The collision cross section varies inversely as the square of the particle energy ($= 2.6 \times 10^{-18}/W^2 \text{ cm}^2$ where W is the particle energy in keV) and is always greater than the fusion cross section, so the majority of nuclei make coulomb collisions before fusion collisions. Even so, it is possible that sufficient fusion collisions can occur for the reactor to be economical. One method of reducing the effect of end loss is to make the reactor very long so that the end area is a small fraction of the total plasma area.

A further method of avoiding end losses is to join the ends forming a plasma loop. This is,

however, unstable since the confining field lines are more closely bunched on the inside of the loop compared to the outside, the pressure is higher on the inside and the plasma drifts outwards. To avoid this, the "Stellarator" experiment uses a figure-eight shaped plasma mixed with magnetic field. Then plasma moving along a field line tends to drift in opposite directions at opposite loops of the figure eight, so if the particle flow around the tube is more rapid than the drift in each loop, the total drift is zero.

Stability. The stability of the plasma/field system is an extremely important factor in the success of a fusion reactor and is the subject of many experiments. A prevalent form of instability is the "flute" or "interchange" in which the plasma develops deep ridges parallel to the field lines. The condition for these ridges to develop is that the field lines are curved with the centre of curvature inside the plasma so that the intensity of the field falls off from the plasma surface. This is a cause of instability in the "z-pinch" and in some regions of the mirror geometry. To prevent this instability, it has been suggested that the field be mixed in the plasma at a skew angle to the confining field. Then, for the instability to develop and the plasma and confining field to be interchanged, the internal skew field has to be stretched and turned, requiring energy; thus, under certain theoretical conditions the plasma should be stable. However experiments with z-pinches show such a system still to be unstable for reasons which are not clear. It may be that finite plasma resistance is affecting the stability criteria.

One method of making the mirror system stable, used by Ioffe, is to add a hexapole field produced by a series of six bars placed parallel to the axis equally spaced around a circumference. Opposing currents flow in alternate bars producing a magnetic field which falls off with distance from the bars. Thus the system has a minimum of magnetic field at its centre and increasing field both radially and towards its mirror ends, and plasma at the center experiences a curved field in all directions with the center of curvature always outside the plasma. This is now recognized to be a particular case of a general class of field geometries called minimum B configurations which are inherently stable against hydromagnetic flute instabilities. Preliminary experiments show a large increase in containment time when the Ioffe bars are in operation.

Another form of flute instability may grow when the plasma is accelerated, as in a rapid magnetic compression or by centrifugal forces if the plasma rotates. These are similar instabilities to the gravitational instabilities of a mercury/water layer, with mercury above, first investigated theoretically by Rayleigh. The origin of the rotation is not definitely understood, but it is known that again the Ioffe bars also suppress these instabilities.

A further and important class of instabilities, the "velocity space" instabilities, can arise for particular conditions of plasma β if the plasma is

not thermalized so that the mean energy parallel to the field lines differs from the energy perpendicular to the lines. When the perpendicular energy is the greater, the plasma tends to form bunches, called "mirror instability," and when the parallel energy is the greater, the plasma wriggles like a fire hose—"fire-hose instability".

These examples of instability are given to illustrate the many types that can occur. It is obvious, however, that the lower the plasma β (so that the vacuum field is approached), especially with a minimum B configuration, the more stable is the plasma.

Conclusions. Only a few problems of fusion research have been mentioned and demonstrated by reference to some of the many experiments proceeding in the field. The main centers of research are Russia (35 per cent of total effort), the United States (25 per cent), Great Britain (10 per cent), and the rest are in Europe. The cost of the research in the United States is about 25×10^6 dollars/yr, in Great Britain, 10×10^6 dollars/yr, in Europe, 6.2×10^6 dollars/yr.

At present (1964), the theta pinch has the greatest nT of 10^{12} at densities of 10^{16} to 10^{17} and temperatures $\approx 10^7$ K. The major difficulties are particle loss from the ends, radiation and plasma rotation. Injected mirror experiments have the highest energy (~ 10 to 100 keV), but only low densities of $\sim 10^{10}$ have been attained because of instabilities. Great hopes are centered around the application of minimum B field configurations to the injection mirror experiments.

J. A. REYNOLDS

References

Glasstone, S., and Lovberg, R. H., "Controlled Thermonuclear Reactions," Princeton, N. J., D. Van Nostrand Co., 1960.

Cross-references: BREMSSTRAHLUNG, FISSION, NUCLEAR REACTIONS, NUCLEAR REACTOR, NUCLEONICS, PLASMAS.

G

GAS LAWS

The term "gas law" refers to the thermodynamic equation of state of a gas, which is an equation relating the pressure p , the volume V , the absolute temperature T , and the number of moles ν . The equation of state is a valid relation when and only when the gas is in a state of thermodynamic equilibrium; the pressure and temperature are then constant and uniform throughout the volume occupied by the gas.

Ideal or Perfect Gas. The ideal or perfect gas is defined thermodynamically by the two conditions: (1) it obeys the equation of state: $pV = \nu RT$ where R is the gas constant per mole ($R = 8.3169 \times 10^7$ erg mole⁻¹ C⁻¹), and (2) the internal energy U is independent of pressure and volume and is a function only of the temperature ($(\partial U/\partial V)_T = 0$). The statistical-mechanical definition of an ideal gas is that it is a gas of noninteracting molecules, i.e., the molecules exert no appreciable forces of attraction or repulsion on each other. Since the notion of a finite "size" of a molecule connotes the existence of a repulsion which prevents two molecules from overlapping each other, the molecules of an ideal gas must be of negligible "size." The two thermodynamic properties can be deduced from the statistical mechanical definition.

The ideal gas equation: $pV = \nu RT$ embodies the experimental laws of Boyle, Charles and Gay-Lussac. It can be derived either from kinetic theory or from statistical mechanics. It is often written in the form: $p = nkT$ where n is the molecular number density and k is Boltzmann's constant ($k = 1.3804 \times 10^{-16}$ erg C⁻¹). In the case of a mixture of inert, ideal gases, each gas obeys the equation: $p_i = n_i kT$ where p_i and n_i are, respectively, the partial pressure and partial density of the i th component gas. Boyle's law will not hold if the gases in the mixture react chemically, since a change in p or V will in general change the value of ν .

Real or Imperfect Gas. The ideal gas law is, of course, only an approximation which holds at temperatures sufficiently far above the critical temperature and at sufficiently low densities. The ordinary properties of bulk matter in the liquid and solid states require the existence of strong intermolecular repulsions which endow the molecules with a finite "size" and also require the existence of attractive forces to hold the molecules

together. The equation of state of a real gas is therefore determined by the nature of the intermolecular forces. One of the earliest, simplest, and most useful equations is that of van der Waals

$$\left(p + \frac{a}{V^2}\right)(V - b) = RT \text{ (for 1 mole)} \quad (1)$$

where a and b are constants, determined empirically for each gas, which are related to the attractive and repulsive forces, respectively. This equation can be related theoretically, in first approximation, to a molecular model in which the molecules are represented by rigid elastic spheres that weakly attract each other. The van der Waals equation accounts qualitatively for the liquid-vapor phase transition. The constants a and b can be determined from critical point data.

Other equations of state for an imperfect gas have been proposed which are more accurate than the van der Waals equation, e.g., the equations of Dieterici, Berthelot, Beattie-Bridgeman, and Benedict-Webb-Rubin. These empirical equations are useful in treating the thermodynamic properties of gases at high densities. At low densities, the empirical equations have been superseded by the virial equation of state

$$\frac{pV}{RT} = 1 + \frac{B(T)}{V} + \frac{C(T)}{V^2} + \frac{D(T)}{V^3} + \dots \text{ (for 1 mole)} \quad (2)$$

where $B(T)$, $C(T)$, and $D(T)$ depend on the nature of the gas and are called the second, third and fourth virial coefficients, respectively. The departures of a gas from ideality are represented in this case by a power series in the density. We may rewrite Eq. (2) as

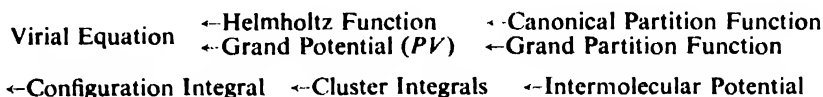
$$p/kT = n + \hat{B}(T)n^2 + \hat{C}(T)n^3 + \hat{D}(T)n^4 + \dots \quad (3)$$

where \hat{B} , \hat{C} , \hat{D} are the virial coefficients referred to one molecule.

The basic experimental problem in this field is to measure the virial coefficients of different gases as functions of the temperature. The higher-order coefficients beyond $B(T)$ and $C(T)$ are very difficult to measure. The basic theoretical problem is to calculate the virial coefficients from an assumed form for the intermolecular potential

energy and, ultimately, to derive the intermolecular potential from quantum mechanics.

Statistical Mechanics of the Imperfect Gas. The derivation of the virial equation of state from the intermolecular potential involves several steps which may be summarized as follows:



The two routes indicated are via the canonical and the grand ensembles. The most difficult step is the evaluation of the configuration integral:

$$Q_N = \int \cdots \int \exp[-\Phi(r_1, r_2, \dots, r_N)/kT] dr_1 dr_2 \cdots dr_N \quad (4)$$

where Φ is the total intermolecular potential energy of the gas of N molecules. The proper way to evaluate Q_N was first sketched by Ursell and later carried through by Mayer who assumed central forces that were pairwise additive, i.e.,

$$\Phi = \sum_{i < j} u(r_{ij})$$

where $u(r_{ij})$ is the potential between molecules i and j . Neither of these assumptions is correct, but they appear to be good approximations, and with their aid, it is possible to evaluate Q_N rigorously in terms of the so-called cluster integrals, b_l , which are integrals over the coordinates of l molecules only. $B(T)$ is obtained directly from b_2 , $C(T)$ is obtained from b_3 and b_2 , and the l th virial coefficient requires evaluation of the cluster integrals up through b_l . Explicit formulas for $\hat{B}(T)$ and $\hat{C}(T)$ are:

$$B(T) = -\frac{1}{2V} \iint f_{12} d\tau_1 d\tau_2 \quad (5)$$

$$\hat{C}(T) = -\frac{1}{3V} \iiint f_{12} f_{23} f_{13} d\tau_1 d\tau_2 d\tau_3 \quad (6)$$

where $f_{ij} \equiv \exp[-u(r_{ij})/kT] - 1$. Higher coefficients in the virial series are increasingly more difficult to evaluate.

The calculations just described are based on the classical Maxwell-Boltzmann statistics and are sufficiently accurate for most gases at ordinary temperatures. However, in the case of the lightest gases H_2 and He and especially at low temperatures, it is necessary to introduce quantum corrections in the equation of state.

Intermolecular Forces and the Equation of State. In order to calculate the cluster integrals and virial coefficients, one must first choose a form for the intermolecular potential $u(r)$. In principle, the potential $u(r)$ is determined by quantum mechanics for any pair of molecules and could be found by solving the Schrödinger equation. In practice, this is virtually impossible, and quantum-mechanical calculations have been made only for the very simplest molecules. In the case of interactions between neutral, nonpolar, spherical

molecules, e.g., noble-gas atoms, the quantum-theoretical interaction energy can be approximately decomposed into several parts, of which the two most important are the *dispersion energy* and the *valence repulsion energy*. The former corresponds to the van der Waals attraction and the

latter to the van der Waals repulsion. The dispersion energy varies inversely with the sixth power of the distance. The valence-repulsion energy takes account of the short-range repulsion that sets in when the electron distributions of the two molecules begin to overlap, and it is associated with the Pauli exclusion principle. There is no simple, general form for the dependence of the valence repulsion potential on the distance; it is often empirically represented by Ae^{-ar} or by μr^{-n} where $n \geq 12$ is commonly used. In the case of molecules that possess permanent electric dipole or quadrupole moments, there are additional contributions to the van der Waals attraction but these are usually less important than the dispersion energy (H_2O is an exception).

In the absence of a complete quantum-mechanical expression for the intermolecular potential, it is necessary to approximate the potential by a semi-empirical formula, containing one or more adjustable constants, which is chosen on the grounds of physical plausibility and mathematical convenience. The semi-empirical force law is then used to calculate macroscopic properties that are known from experiment, and the parameters in the force law are adjusted to give the best agreement with experiment. Given the form of the intermolecular potential, it is possible to calculate not only the virial coefficients in the equation of state but also the kinetic-theory transport coefficients (i.e., the viscosity coefficient, the thermal conductivity, and the various diffusion coefficients of the gas) and the density, compressibility, and sublimation energy of the solid. A particular functional representation of the intermolecular potential can be considered satisfactory only if it is possible to secure agreement with all experimental data involving a particular pair of molecules with a single choice of the parameters that appear in the law of force.

The semiempirical law that is most frequently used to represent the interaction between nonpolar molecules is the *Lennard-Jones* (12, 6) potential:

$$u(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^6] \quad (7)$$

where ϵ and σ are parameters characterizing the particular pair of molecules. This simple two-parameter function, when inserted in Eq. (5), predicts a temperature variation of the second virial coefficient in good agreement with experiment. The same potential also explains the temperature variation of the viscosity coefficient over

a substantial temperature range, but the agreement is poor for the highest experimental temperatures. Third virial coefficients calculated from Eq. (6) do not agree with experiment at the lowest temperatures. Calculations of virial coefficients and transport coefficients of gases and of equilibrium properties of the crystal lattices have also been made for other simple semiempirical potential functions, but the results are not very different from those found with the (12, 6) potential.

Further advances in this field will come, on the theoretical side, from a more detailed knowledge of the intermolecular forces, and on the experimental side, from more accurate ways of measuring the virial coefficients.

Dense Gases. High-density gases cannot be conveniently represented by the virial equation of state because of the slow convergence of the virial series. Furthermore, the theoretical evaluation of the higher virial coefficients on the basis of any plausible molecular model would meet with great computational difficulties. Other approaches are therefore needed, e.g., the empirical equations of state already mentioned and the principle of corresponding states. In the latter method, one introduces the reduced, dimensionless variables: p/p_c , $V_r = V/V_c$, and T/T_c where the subscript c refers to the critical point. The principle of corresponding states then asserts that all substances obey the same equation of state in terms of the reduced variables. The variables may also be reduced in terms of intermolecular potential parameters.

A promising theoretical approach to the equation of state of a dense gas or liquid is provided by the method of the radial distribution function $g(r, n, T)$. Because of intermolecular forces, the actual density at a small distance r from a given molecule is different from the bulk density n and is represented by $ng(r, n, T)$. Thus the radial distribution function measures the effect of intermolecular forces on the probability of finding two molecules close together. While it is difficult to determine $g(r, n, T)$ theoretically, it can be found experimentally from the diffraction pattern observed when x-rays are scattered by the fluid.

•

R. D. PRESENT

References

- Cowling, T. G., "Molecules in Motion," London, Hutchinson & Co., Ltd., 1950 (for the general reader).
 Hill, T. L., "Statistical Mechanics," Chs. 5 and 6, New York, McGraw-Hill Book Co., 1946 (statistical mechanics of imperfect gases and dense fluids).
 Hirschfelder, J. O., Curtiss, C. F., and Bird, R. B., "Molecular Theory of Gases and Liquids," Chs. 3 and 4, New York, John Wiley & Sons, 1954 (covers all aspects of the subject and is the standard reference in this field).
 Present, R. D., "Kinetic Theory of Gases," Ch. 6 and 12, New York, McGraw-Hill Book Co., 1958

(kinetic theory of the second virial coefficient; intermolecular forces).

Rushbrooke, G. S., "Introduction to Statistical Mechanics," Ch. 16, London, Oxford University Press, 1949 (good introduction to imperfect-gas theory).

Cross-references: COMPRESSIBILITY, GAS; GASES: THERMODYNAMIC PROPERTIES; KINETIC THEORY; THERMODYNAMICS.

GASES: THERMODYNAMIC PROPERTIES

Fundamental Principles. The thermodynamic properties of a substance may be classified as either reference properties, energy functions, or derived properties.¹ The reference properties of a single component system are pressure, volume, temperature, and entropy. For a specific amount of a pure gas, it is necessary to specify only two of these reference properties to fix the state of the system and its properties. For mixtures of gases the composition must also be specified to completely fix the system. The energy functions are internal energy, U ; enthalpy, H ; Helmholtz free energy, A ; and Gibbs free energy, G . These functions represent the energy available for performing useful work under various process conditions. Derived properties include specific heat, fugacity, activity coefficient, compressibility factor, and the Joule-Thomson expansion coefficient.

Properties are termed intensive if they are independent of the amount of the material. Examples are pressure and temperature. Properties such as volume and entropy, which are dependent on the amount of material, are termed extensive.

Absolute values may be determined for the reference properties, but the energy functions must be determined relative to an arbitrary zero reference point. The internal energy, U_0 , of the ideal gas at the absolute zero of temperature is generally taken as the zero reference point of the enthalpy and free energy functions. Other reference points include a zero value for the enthalpy of the ideal gas at the ice point, $H_{273.15}$, and another in which the sensible enthalpies are combined with chemical energies.² In the latter base, the value of $H_{298.15}$ is zero for the assigned reference elements so that the values of $H_{298.15}$ for the various compounds are equal to their heats of formation from the assigned reference elements.

Thermodynamic properties of gases are calculated for both the ideal gas state and the real gas state. A gas is defined to be ideal if it follows the simple equation of state, $PV = RT$. Gases behave in this manner only at very low pressure, but the ideal gas state is a convenient reference state for the calculation of the thermodynamic properties. Thus, the thermodynamic standard state³ is defined as the ideal gas at unit pressure at each temperature, and it is denoted by a superscript degree mark as in H° and S° . The ideal gas properties have been calculated for many substances, but the real gas properties are known for relatively few substances.

Thermodynamic properties are used in the calculation of energy balances, reaction compositions at chemical equilibrium, reaction temperatures, and the work involved in the compression or expansion of gases in various systems.

Ideal Gas Properties. The thermodynamic properties of an ideal gas such as C_p° , $(H^\circ - H_0^\circ)/T$, S° and $(G^\circ - G_0^\circ)/T$ are calculated from theoretical equations and from an analysis of spectroscopic and molecular structure data.⁴ These complex calculations are based upon the contributions from all of the energy states available to the molecule, such as translational, electronic, vibrational, and rotational. Contributions from excited electronic states are important for diatomic molecules at higher temperatures but are entirely negligible for most polyatomic molecules. Vibrational energy levels are obtained from an analysis of infrared and Raman spectroscopic data by applying the principles of wave mechanics and group theory.^{4,5} The interpretation of the spectra includes the assumption of a model for the molecule with parameters such as bond lengths, bond angles, and force constants. The parameters are varied within certain limits until the best agreement with observed spectra is obtained. Rotational energy levels are observed in infrared, Raman, and microwave spectra. The rotational energy includes not only the rotation of the molecule as a whole but also internal rotations by groups of atoms within the molecule and a pseudorotation in some ring molecules.⁶

For higher orders of accuracy in the calculation of the thermodynamic properties, additional contributions may be determined that are caused by the interaction between vibrational and rotational motions, centrifugal distortion of the molecule during rotation, and anharmonicity of the vibrations. Another contribution due to nuclear spin can be, and generally is, neglected for all molecules except H_2 and D_2 since it causes a detectable effect on measurable quantities only at very low temperatures. Adjustments are made for the presence of isotopes in some diatomic and polyatomic molecules.

The calculation of the thermodynamic properties of an ideal gas is based not only on theory but also on accurate experimental vapor heat capacity, heat of vaporization, and low-temperature calorimetric data.⁶ The ideal gas heat capacity C_p° and entropy S° are derived from these data and are compared to theoretical values. When differences are found, the theoretical calculations are revised until there is good agreement. In this manner, new information is gained about the conformation of the molecule, its frequencies of vibration, etc. Thus, experimental data provide a firm base to test theoretical calculations and improve the calculation of all of the thermodynamic properties.

Theoretical calculations become increasingly complex as the molecular size increases. Thus, a method of increments has been devised to calculate the thermodynamic properties of large molecules based on an "anchor compound" for a given homologous series.^{1,7}

Real Gas Properties. The thermodynamic properties of real gases are determined primarily from experimental compressibility (pressure-volume-temperature) measurements (see COMPRESSIBILITY, GAS). All other properties are calculated either from equations of state or from a correlation of the individual experimental data points. In addition, the properties may be estimated from generalized correlations of the compressibility factor ($Z = PV/RT$).

Experimental compressibility measurements have been made by a variety of methods⁸⁻¹⁰ such as constant volume cells (Euken and Meyer), variable volume cells (Beattie and Douslin, Michels, and Sage and Lacey), expansion systems with variable sample mass (Burnett), and differential systems (Whytlaw-Gray). The apparatus of Beattie and Douslin is used to measure both isometrics and isotherms up to 350°C and 400 atmospheres. The Michels apparatus may be used up to 3000 atmospheres but is limited to a temperature of 150°C. Sage and Lacey have made extensive measurements on both gas and liquid systems up to 240°C and 670 atmospheres.

Equations of state have been derived from compressibility data to represent the behavior of a gas over wide ranges of temperature and pressure.^{8,11} Numerous equations have been published but one of the most important is the virial equation,

$$PV = RT(1 + B_1V + C_1V^2 + D_1V^3 + \dots) \quad (1)$$

It is quite important because the parameters B , C , D , etc., are related to the interactions between molecules according to the intermolecular potential energy theory. Other equations having wide applications are those of Beattie and Bridgeman and of Benedict, Webb, and Rubin (see GAS LAWS).

The energy functions, entropy, and heat capacity of a real gas are calculated as the sum of the ideal gas properties and a correction for the non-ideality of the gas. The corrections for the non-ideality of the gas are called difference or departure functions. For example, $S - S^\circ$ is the entropy of the gas in the real state less that of the gas in the standard state at the same temperature. Theoretical equations needed for the computation of the thermodynamic properties have been derived in terms of pressure, volume, temperature, and the first derivatives $(\partial V/\partial T)_p$ or $(\partial P/\partial T)_v$.¹¹

For example:

$$H - H^\circ = \int_v^\infty [P - T(\partial P/\partial T)_v]dv + PV - RT \quad (2)$$

The heat capacity differences $C_p - C_p^\circ$ and $C_v - C_v^\circ$ are functions also of the second derivatives $(\partial^2 V/\partial T^2)_p$ or $(\partial^2 P/\partial T^2)_v$ depending upon whether the equations are written in terms of P and T , or V and T . The quantities appearing in Eq. (2) are usually evaluated from equations of state. However, the most accurate properties are those which are calculated from an analysis of

isometric and isothermal P - V - T data.¹² The slopes of the isometrics $(\partial P/\partial T)_r$ are found by analytical, residual, and graphical techniques. The integrals as in Eq. (2) are integrated graphically or numerically.

Extensive correlations of data have been made to develop methods for estimating the properties of gases.^{1,11} One method based on the theory of corresponding states presents the thermodynamic properties as a function of reduced temperature ($T_r = T/T_c$), reduced pressure ($P_r = P/P_c$), and the compressibility factor, Z_c . The subscript c refers to the critical state.

Gas Mixtures. The thermodynamic properties of a mixture of gases may be calculated, but the procedures are only approximate unless compressibility data are available for the particular mixture.^{8,11} Since few data for mixtures are available, the properties must be estimated from: (a) the equations of state of the pure gases assuming either additive volumes or additive pressures, (b) an equation of state for the mixture, or (c) generalized correlations of the compressibility factor based on pseudoreduced conditions.¹

ROLAND H. HARRISON

References

1. Hougen, O. A., Watson, K. M., and Ragatz, R. A., "Chemical Process Principles Part Two: Thermodynamics," New York, John Wiley and Sons, Inc., 1959.
2. McBride, B. J., Helmel, S., Ehlers, J. G., and Gordon, S., "Thermodynamic Properties to 6000 K for 210 Substances Involving the First 18 Elements," NASA, SP-3001, Washington, D.C., 1963.
3. Canjar, L. N., and Rossini, F. D., "Work of the American Petroleum Institute Research Project 44 on P-V-T Properties," Conference on Thermodynamic and Transport Properties of Fluids, sponsored jointly by British Institute of Mechanical Engineers and International Union of Pure and Applied Chemistry, London, England, July 1957.
4. Herzberg, G., "Molecular Spectra and Molecular Structure II. Infrared and Raman Spectra of Polyatomic Molecules," Princeton, N.J., D. Van Nostrand Co., Inc., 1945.
5. Wilson, E. B., Jr., Decius, J. C., and Cross, P. C., "Molecular Vibrations—The Theory of Infrared and Raman Vibrational Spectra," New York, McGraw-Hill Book Co., Inc., 1955.
6. McCullough, J. P., Pennington, R. E., Smith, J. C., Hossenlopp, I. A., and Waddington, G., "Thermodynamics of Cyclopentane, Methylcyclopentane, and 1, *cis*-3-Dimethylcyclopentane: Verification of the Concept of Pseudorotation," *J. Am. Chem. Soc.*, **81**, 5880 (1959).
7. Scott, D. W., and McCullough, J. P., "The Chemical Thermodynamic Properties of Hydrocarbons and Related Substances," *U.S. Bur. Mines Bull.* 595 (1961).
8. Rowlinson, J. S., in Flugge, S., Ed., "Encyclopedia of Physics," Vol. XII, pp. 1-72, Berlin, Springer-Verlag, 1958.
9. Beattie, J. A., "The Apparatus and Method Used for the Measurement of the Compressibility of Several Gases in the Range 0°-325°C," *Proc. Am. Acad. Arts Sci.*, **69**, 389-405 (1935).
10. Douslin, D. R., Harrison, R. H., Moore, R. T., and McCullough, J. P., "P-V-T Relations for Methane," *J. Chem. Eng. Data*, **9**, No. 3, 358-363 (1964).
11. Beattie, J. A., and Stockmayer, W. H., in Taylor, H. S., and Glasstone, S., Eds., "Treatise on Physical Chemistry," Vol. 2, pp. 187-290, Princeton, N. J., D. Van Nostrand, Co., Inc., 1951.
12. Harrison, R. H., and Douslin, D. R., "Perfluorocyclobutane: The Thermodynamic Properties of the Real Gas," *U.S. Bur. Mines Rept. Invest.* 6475 (1964), 14 pp.

Cross-references: COMPRESSIBILITY, GAS; GAS LAWS; HEAT CAPACITY; KINETIC THEORY; THERMODYNAMICS.

GEODESY

Geodesy comprises the determination of the earth's external form and gravitational field, and the location of points with respect to earth-fixed reference systems. The earth's external form is customarily defined by the geoid: the equipotential of the earth's gravity field which most closely approximates the mean sea level.

The geoid is irregular in form, so that the mathematical representation thereof is necessarily an approximation. The most important approximation is an oblate ellipsoid of revolution, which is conventionally defined by its equatorial radius, a , and the flattening, f , equal to $(a - b)/a$, where b is the polar semidiameter. Location is conventionally expressed in coordinates referred to such an ellipsoid, in terms of the latitude ϕ , the angle between the normal to the ellipsoid and the equator; the longitude λ from the reference meridian, Greenwich; and altitude h above or below the ellipsoid.

If the ellipsoid is considered to be rotating with rate ω , and to be an equipotential for the combined effects (called gravity) of centrifugal and gravitational acceleration, additional parameters customarily required are γ_e , the acceleration of gravity at the equator; and m , the ratio of the centrifugal acceleration at the equator, $\omega^2 a$, to γ_e . γ_e and m are connected to the total mass M contained in the ellipsoid by:

$$kM = a^2 \gamma_e [1 - f + 3m/2 - 15mf/14 + O(f^3)] \quad (1)$$

where k is the constant of GRAVITATION ($6.664 \times 10^{-8} \text{ cm}^3 \text{ g}^{-1} \text{ sec}^{-2}$). The customary formula for the acceleration of gravity γ at geodetic latitude ϕ is:

$$\gamma = \gamma_e [1 + (5m/2 - f - 17mf/14) \sin^2 \phi + (f^2/8 - 5mf/8) \sin^2 2\phi + O(f^3)] \quad (2)$$

The customary formula for the gravitational potential external to the ellipsoid is:

$$V = \frac{kM}{r} \left[1 - J_2 \left(\frac{a}{r} \right)^2 P_2(\sin \phi) - J_4 \left(\frac{a}{r} \right)^4 P_4(\sin \phi) - O(f^4) \right] \quad (3)$$

In Eq. (3), V is written as positive, which is the convention of astronomy and geodesy, contrary to that of physics. In Eq. (3), P_2 and P_4 are Legendre Polynomials, and

$$J_2 = 2f(1 - f/2)/3 - m(1 - 3m/2 - 2f/7)/3 + O(f^3) \quad (4)$$

while J_4 is usually taken as a quantity determined observationally from satellite orbits.

The discrepancies in location of the actual geoid and a well-fitting ellipsoid are nearly always 10^{-5} or less of the radius vector, while the discrepancies in intensity of the gravitational acceleration from that of the standard ellipsoid are nearly always 10^{-4} or less of the total intensity. The mathematical representation of these discrepancies may either be in the form of spherical harmonic coefficients or in the form of mean values for areas; the former being preferable for effects on satellites in orbit, and the latter for use of terrestrial data. The potential theory dealing with the relationship between variations in the location of the geoid and the intensity of gravitational acceleration is known in geodesy as Stokes' theorem.

There are five principal systems of measurement in geodesy.

(1) Horizontal control comprises the determination of the horizontal components of position—latitude and longitude—starting from fixed values for a certain point. It includes measurement of distances over the ground by metal tapes or by pulsing or modulating radio or light signals, and measurement of angles about a vertical axis by theodolites. Over the land, the relative horizontal position of points is obtained either by triangulation—a system of overlapping triangles with nearly all angles measured, but only occasional distances measured; or by traverse—a series of measured distances at measured angles with respect to each other; or by trilateration—a system of overlapping triangles with all sides measured. Much of the land area of the world is covered by triangulation, which gives the difference in latitude and longitude between points in the same network with a relative error of about 10^{-5} .

(2) Vertical control comprises the determination of heights, which is performed separately from horizontal control because of irregularities in atmospheric refraction. The most accurate method, leveling, measures successive differences of elevation on vertical staffs by horizontal lines of sight taken at intermediate points over short distances (less than 150 meters) balanced so as to minimize differential refraction effect. The datum to which vertical control refers is mean sea level as determined by tide gages. The accuracy is such

that the error in difference of elevation between points on the same principal network should be a few tens of centimeters or less.

(3) Geodetic astronomy comprises the determination of the direction of the gravity vector and the direction of the north pole at a point on the ground. Astronomic longitude is the angle between the meridian of the gravity vector and the Greenwich meridian and is determined by measuring the time of intersection of a line of sight by a star. Astronomic latitude is the angle between the gravity vector and the equatorial plane, and is determined by measuring the maximum altitude attained by a star. In these types of astronomic observation, several stars are normally observed which are selected so as to minimize error due to atmospheric refraction. Astronomic azimuth is determined by the measurement of the horizontal angle between a target and Polaris or other reference star.

(4) Gravimetry comprises the determination of intensity of gravitational acceleration. Most gravimetric observations are made differentially, by determining the change, with change in location, of the tension on a spring supporting a constant mass. These measurements are connected through a system of reference stations to a few laboratory determinations of absolute acceleration of gravity. The relative accuracy of gravimetry is about $\pm .001 \text{ cm sec}^{-2}$ on land and $\pm .005 \text{ cm sec}^{-2}$ at sea. The principal difficulty in its geodetic application is irregular distribution of observations.

(5) Satellite tracking comprises the determination of the directions, ranges, or range rates of earth satellites from ground fixed stations. These observations will be affected both by errors in positions of the station with respect to the earth's center of mass and by perturbations of the orbit by the earth's gravitational field; hence, in conjunction with a suitable dynamical theory for the orbit, they are used to determine the position of tracking stations and the variations of the gravitational field. To minimize refraction effect, directions are determined by photographs of the satellite against the background of fixed stars. Satellites can also be used as elevated targets by simultaneous observations from several ground stations.

The principal practical application of geodesy is to provide a distribution of accurately measured points to which to refer mapping, navigation aids, engineering surveys, geophysical surveys, etc. The principle scientific interest in geodesy is the indication of the earth's internal structure by the variations of the gravity field.

Numerical values of the leading geodetic parameters are given in Table 1.

WILLIAM M. KAULA

References

- Bomford, G., "Geodesy," Second edition, 561 pp. London, Oxford University Press, 1962.

TABLE 1. GEODETIC PARAMETERS

Parameter	Standard Value	Current Estimate and Standard Deviation
Mean sidereal rotation rate, ω	$.7292115085 \times 10^{-4} \text{ sec}^{-1}$	$.7292115085 \times 10^{-4} \text{ sec}^{-1}$
Equatorial gravity, γ_e	9.780490 m/sec ²	$9.780306 \pm .000013 \text{ m/sec}^2$
Equatorial radius, a	6378388 meters	$6378160 \pm 15 \text{ meters}$
Flattening, f	1/297.00	$1/298.25 \pm .03$

Heiskanen, W. A., and Vening Meinesz, F. A., "The Earth and Its Gravity Field," New York, McGraw-Hill Book Co., 470 pp., 1958.

Jeffreys, H., "The Earth," Fourth edition, 420 pp., London, Cambridge University Press, 1959.

Kaula, W. M., "Determination of the Earth's Gravitational Field," *Revs. Geophys.*, 1, 507-551 (1963).

O'Keefe, J. A., "Geodesy," New York, Prentice-Hall, 1964.

Veis, G., Ed., "The Use of Artificial Satellites for Geodesy," Amsterdam, North Holland Publ. Co., 1963.

Cross-references: \approx TRODYNAMICS, ASTRONOMY, ASTRONAUTICS, GEOPHYSICS, GRAVITATION, POTENTIAL, ROTATION—CIRCULAR MOTION.

GEOPHYSICS

Geophysics is the physics of the earth and the space immediately surrounding it and the interactions between the earth and extraterrestrial forces and phenomena. It consists of a number of interlocking sciences dealing with physical properties of the earth, its interior and atmosphere, its age, motions and paroxysms, and their practical applications. All of these sciences use the methods of physics for measurements and analysis. From observational material, often of an indirect nature, attempts are made to derive abstract models of states and processes through advanced mathematical concepts and, in some cases, through statistical relations.

Geophysics is an ancient science. In its early stages it was developed by the Greeks who attempted to determine the shape and size of the Earth (Eratosthenes, 275–194 B.C.). Among its most illustrious contributors have been Galileo Galilei (1564–1642); Sir Isaac Newton (1642–1727) who dealt with the motions of the earth and its gravitational field; Karl Friedrich Gauss (1777–1855), who developed the theory of the magnetic field; and Vilhelm Bjerknes (1862–1951) who laid the foundation for the hydrodynamic theories of the atmosphere and the oceans. During the current century, its roster of distinguished scientists includes: L. Vegard (polar aurora); Sidney Chapman (aeronomy); C.-G. Rossby (meteorology); H. U. Sverdrup (oceanography); Sir Harold Jeffreys; F. A. Vening-Meinesz (structure of the earth); and B. Gutenberg and J. B. Macelwane (seismology).

A major milestone in the development of the science was the International Geophysical Year (IGY), followed by the International Geophysical Cooperation, from 1957–1959, when over 8000 scientists of 66 nations collaborated in a coordinated attack on the remaining physical mysteries of our planet. The IGY gave birth to man's most spectacular ventures to date: the launching of artificial earth satellites and the conquest of the icy wastes of the Antarctic continent in the quest for solutions of important geophysical problems. IGY took place during an interval of high solar activity and was followed in 1964–65 by the year of the quiet sun (IQSY) to cover the whole range of solar influences on the earth (see INTERNATIONAL GEOPHYSICAL YEAR AND INTERNATIONAL YEARS OF THE QUIET SUN).

A brief survey of the various subfields of geophysics follows:

Geochronology is the study of the age of the earth and its various geological formations. Inadequate earlier methods of sedimentology have been replaced by the use of radioactive decay constants and isotope ratios. This has made it possible to date approximately all major geological eras. For the oldest rocks, the decay of U^{238} to Pb^{206} has led to ages of around 3×10^9 years.

GEODESY deals with the size and shape of the earth and its gravitational field. Because of the rotation of the earth, its lack of absolute rigidity, crustal mass distribution, and tidal forces, the shape is not spherical but approximates that of a triaxial ellipsoid. The polar diameter is shorter than the equatorial. This polar flattening has been determined by measurements on the surface, lunar observations and, since, 1958 satellite motions. The best estimate of the flattening is 1:298. The actual figure of the earth, which is irregular, is referred to as the *geoid*, which represents an isopotential surface. The undulations of the geoid have been determined by gravity measurements. The measured values of gravity depend on latitude because of the flattening and the variation of the centrifugal force from pole to equator. The normal value of gravity at the earth's surface in centimeters per second per second, is represented by

$$\gamma = 978.0516[1 + 0.005291 \sin^2\phi - 0.0000057 \sin^2 2\phi + 0.0000106 \cos^2\phi \cos 2(\lambda + 6^\circ)]$$

where ϕ and λ are latitude and longitude, respectively.

Gravity measurements have shown that in spite of large mass differences at the surface, the earth is nearly in isostatic equilibrium. Various crustal blocks act as if they floated in a dense subcrustal material. The undulations of the geoid do not exceed 80 meters. Approximate dimensional figures for the earth are: surface area $510.1 \times 10^6 \text{ km}^2$; volume $1,083 \times 10^9 \text{ km}^3$; average density 5.517 g/cm^3 ; mass 5.975×10^{27} grams; equatorial radius 6,378.163 km.

The deformations of the solid earth by tidal forces form a specialty. The twice daily occurring *tides* are observed by deflections of the vertical or variations of gravity. For the lunar tide, the variations amount to about 0.168 milligal, for the sun to 0.075 milligal. (One milligal equals 10 microns per sec per sec.) The maximal elevation of the geoid is 36 cm, the largest depression 18 cm, for the lunar effect; the total solar tide can reach 25 cm. The combined total at new and full moon is 79 cm.

Geomagnetism deals with one of the most important physical characteristics of the earth. Its magnetic field is about 1/2 gauss in strength. It consists of an internal part, about 9/10 of the total and an external part of 1/10. The internal part is assumed to originate from electromagnetic fields caused by differential rotation between the earth's core and its mantle. The external part is caused by an ionospheric ring current. The field can be represented by a dipole, which fluctuates in direction both in time and space. Short-periodic fluctuations are brought about by solar disturbances which cause invasions of protons and electrons into the high atmosphere. These disturbances, which follow broadly solar activity as expressed by sunspots, cause minor and major fluctuations of the earth's magnetic field. The major disturbances, called magnetic storms, can be observed worldwide. The same particle bombardments cause the *polar aurora* through excitation of oxygen, hydrogen, and nitrogen atoms in the high atmosphere. Much slower variations of the magnetic field are tied to the internal field and possibly connected with convective currents. Paleomagnetic evidence indicates that the magnetic poles are not fixed and that the field has reversed direction several times during the earth's history. In the present era, the magnetic poles and poles of rotation do not coincide.

It has been hypothesized that the migration of the magnetic poles is also an indication of migrations of the inertial axis of the earth. The distribution of the field over the surface of the earth has been mapped in form of the total intensity, the horizontal and vertical intensity, the inclination and declination. The latter is of great practical importance for use of magnetic compasses in direction finding. The influence of the earth's magnetic field extends outward into space for many earth radii. It influences the path of invading particles and is responsible for the configuration of the van Allen belts.

SEISMOLOGY deals with the study of earth-

quakes. Most of them can be attributed to breaks in the earth's crust. A few may be subcrustal and caused by phase transformations. Some minor ones are associated with collapse of cavities and volcanic eruptions. All of them are characterized by sudden release of elastic waves. These are longitudinal, transverse, Rayleigh or Love (surface) waves. Through analysis of their travel time from the point of origin of receiving seismograph stations, they permit deduction of the layering and elastic constants of the interior of the earth. Like other waves, earthquake waves can be reflected and refracted upon entering a different medium. The internal constitution of the earth has been derived primarily from seismic evidence. This reveals an upper and lower crust, an upper and lower mantle, and the core. The boundary between crust and mantle was first discovered by A. Mohorovičić in 1909. The depth is variable. In some oceanic areas, it may be only 5 km; in continental areas, it is usually 30 to 40 km and the maximum is estimated at 60 km. The quest for knowledge about the petrographic constitution of the mantle has led to the project attempting to drill in one of the suboceanic thin crustal areas to the Mohorovičić discontinuity, the so-called Mohole. The boundary of the core is about 2900 km deep. The density in the mantle is from 3 to 6 g/cm³, that of the core from 10 to 17 g/cm³. The temperature of the interior has been variously estimated at 1500 to 4000 C. Earthquakes are primarily tied to areas of geologically recent mountain formation, the great island arcs, and the adjacent great rifts. The same areas usually show also the greatest contemporary volcanic activity.

Volcanology deals with some of the most spectacular phenomena in geophysics. They include explosive eruptions, lava flows, gaseous exhalations, magma intrusions, geysers and hot springs. Study of the earth's active volcanoes revealed that release of tectonic pressure may lead to volcanic eruptions, but no reliable systems of forecasting impending volcanic activity have yet been devised. The greatest eruption within human memory was that of Krakatoa on May 20, 1883 which threw about 4 km³ dust into the air up to stratospheric levels. This dust stayed in suspension for over two years, interfering with the earth's heat balance. Some hypotheses of climatic fluctuations are based on the varying intensity of eruptive volcanism during the earth's history.

METEOROLOGY and *Aeronomy* are concerned with the physical state and the motions of the atmosphere, which is divided into a number of layers. The lowest is the troposphere with an average thickness of 7 to 8 km in polar regions and 13 km in the equatorial zone. Temperatures decrease to the interface, called tropopause, with the next layer the stratosphere. At the tropopause, polar temperatures average around -55°C , in equatorial regions -80°C . In the stratosphere, temperatures stay nearly isothermal with height and increase again above 25 km. Above the stratosphere are the mesosphere and ionosphere, and the outermost layer, the exosphere, gradually fades into the

plasma continuum between earth and sun. In these higher layers of the atmosphere, complex interactions between the fluxes of electromagnetic radiation of various wavelengths and corpuscular radiation from the sun on one side and the low-density concentrations of atmospheric gases on the other side take place. The particulate radiations are also governed by the earth's magnetic field. Radiations of short wavelength cause a variety of photochemical reactions, the most notable of which is the creation of a layer of ozone acting as an effective absorber of solar ultraviolet and thus causing a warm layer at 30 km in the atmosphere. The upper atmosphere as an absorber of primary cosmic rays shows many interesting nuclear reactions and is an important natural source of radioactive substances including tritium and carbon 14 which are used as tracers of atmospheric motions and as criteria of age.

Most manifestations of weather take place in the troposphere. They are governed by the general atmospheric circulation which is stimulated by the differential heating between tropical and polar zones. The resulting motions in the air are subject to the laws of fluid dynamics on a rotating sphere with friction. They are characterized by turbulence of varying time and space scale. Evaporation of water from the ocean and its transformation through the vapor state to droplets and ice crystals, forming clouds and precipitation, are important symptoms of the weather-producing forces (see METEOROLOGY).

Hydrology studies the water cycle on the earth in detail. It includes the runoff from precipitation, the surface courses of water and their floods, the deposited forms of water as snow and ice, and the return of water to storage underground or the ocean. The study of glaciers and their mass budget is an important phase of this field and contributes to the as yet speculative hypotheses about ice ages.

Oceanography has a broad overlap with both meteorology and hydrology. It includes the study of wind-driven waves and currents, and the storage and release of heat to the atmosphere. The tidal motions and their dependence upon configuration of ocean basins and coastal lines were among the earliest geophysical phenomena observed and analyzed by man. They were also the first to be predicted by a computer. The density differences of ocean waters, predicated on temperature and salinity, cause three-dimensional internal circulations in the deeper ocean layers. Much of this remains to be observed and explained. Of great practical importance to inhabitants of shore lines are the *tsunamis* which are gravity waves caused by underwater earthquakes. Because their movement is much slower than that of elastic seismic waves, warnings can be issued through prompt evaluation of seismic records.

Geophysical Prospecting is based on physical methods derived from the study of earthquakes and magnetic and gravitational fields. These procedures, highly refined, are being used in the discovery of mineral resources. Inhomogeneities

in the crust caused by inclusions of different physical characteristics can be noted from the surface by local anomalies of the magnetic or the gravitational field. Miniature earthquakes produced by explosions will disclose layering, salt domes and other geological features. Distortion of artificially created electric fields also can lead to ore bodies or similar anomalies. Interpretation of radioactivity observed at the surface or in boreholes also permits geological deductions. Extremely sensitive apparatus, especially for magnetometric mapping, can be used from the air and permits rapid surveying of vast areas.

H. E. LANDSBERG

References

- Chamberlain, J. W., "Physics of the Aurora and Airglow," New York, Academic Press, 1961.
 Flugge, S., Ed., "Handbuch der Physik" (Encyclopedia of Physics); Bartels, J., group editor, "Geophysics," Vols. 47 and 48, Berlin, Springer, 1956 and 1957.
 Geophysics Research Directorate, Air Research and Development Command, U.S. Air Force, "Handbook of Geophysics," revised edition, New York, The Macmillan Co., 1960.
 Jacobs, J. A., Russel, R. D., and Wilson, J. Tuzo, "Physics and Geology," New York, McGraw-Hill Book Co., 1959.
 Jeffreys, Harold, "The Earth," Fourth edition, Cambridge, Cambridge University Press, 1959.
 Landsberg, H. E., and van Mieghem, J., Eds., *Advan. Geophys.*, 1-10 (1952-1964).

Cross-references: FLUID DYNAMICS, GEODESY, GRAVITATION, INTERNATIONAL GEOPHYSICAL YEAR AND INTERNATIONAL YEARS OF THE QUIET SUN, MAGNETISM, METEOROLOGY, PLANETARY ATMOSPHERES, SEISMOLOGY.

GRAVITATION

Gravitation is the phenomenon characterized by the mutual attraction of any two physical bodies. This universal character of the gravitational force was first recognized by Sir Isaac Newton who also gave its quantitative expression. For point masses or spherical bodies a simple expression results:

$$F = \frac{GM_1M_2}{R^2} \quad (1)$$

In addition to the masses M_1 , M_2 of the two bodies and their distance apart R , the force depends only on a constant $G = 6.670 \times 10^{-8}$ dyne cm² gm⁻² which is independent of all properties of the particular bodies involved. The same force law describes the motion of the planets around the sun, of the moon around the earth, as well as the falling of an apple to the earth. A body moving under an inverse square law as given in Eq. (1) satisfies the three laws established by Kepler for the motion of the planets around the sun:

(1) The planets move in elliptical orbits with the sun at one focus (the general orbit is a conic section) (Fig. 1).

(2) The radius vector sweeps out equal areas in equal times.

(3) The square of the period of revolution is proportional to the cube of the semi-major axis: $a^3 = (2\pi)^{-2} GM_{\odot} T^2$. Here M_{\odot} is the mass of the sun and T is the period of the planet.

These results together with a detailed analysis of anomalies in the motion of the moon established the correctness of the Newtonian theory of gravitation (see KEPLER'S LAWS).

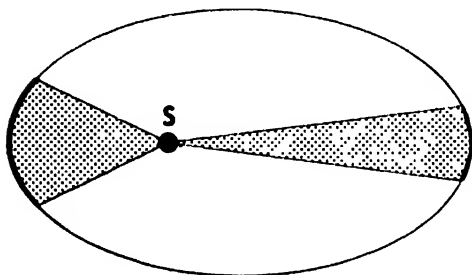


FIG. 1. An elliptical orbit for a planet around the Sun. The shaded areas indicate equal areas swept out in equal times at different parts of the orbit. Clearly, the speed of the planet varies with its position in its orbit.

Recently, careful calculations have been carried out to determine the orbits of the artificial satellites which have been launched by the United States and the Soviet Union. These have required modifications in the force law Eq. (1) to take into account the deviation of the earth's figure from a sphere and the anisotropy of the earth's density as well as the atmospheric drag. The success of the space program to date is an additional tribute to Newton's genius. Other calculations study powered space flight in order to examine possible orbits for exploration of the Solar System. There is every reason to believe that Newton's gravitational theory is sufficiently accurate for this purpose. Einstein's modification of the theory, to be described below, will probably have little effect on our space program for some time to come.

The *weight* of a body of mass M on the earth is the force with which it is attracted to the center of the earth. On the surface of the earth the weight is given by

$$W = Mg$$

where the *acceleration due to gravity* is obtained from Eq. (1):

$$g = \frac{GM_E}{R_E^2} = 980 \text{ cm/sec}^2 = 32 \text{ ft/sec}^2$$

All freely falling bodies near the surface of the earth are accelerated at the same rate g . It is for this reason that Galileo found that both light

and heavy objects take the same time to reach the ground when dropped from the Leaning Tower of Pisa.

An astronaut is said to be in a state of *weightlessness* when he is in orbit. Strictly speaking, he still has weight for the earth's gravity still acts on him. Otherwise he would fly off into outer space. However, when in free fall, the local effects of the gravitational field are eliminated for the astronaut. Objects which are released fall together with him and hence remain in his vicinity unlike the situation on the ground. Therefore, the organs of the body respond as though the gravitational field were absent and this gives the sensation of weightlessness.

Gravitational Field. According to Newtonian theory, the sun exerts the gravitational force directly on the earth without an intervening medium for transmitting that force. The behavior of such forces is called "action at a distance." To overcome the conceptual difficulty of a force acting directly over large distances, one assumes that a *gravitational field* fills all space. The force acting on any mass is determined by the gravitational field in its neighborhood. Thus, at the point P a distance R from the center of the earth, the gravitational field has the magnitude

$$\mathcal{G} = \frac{GM_E}{R^2}$$

and magnitude of the force on a mass M at P is simply $F = M\mathcal{G}$. Note that the field is to exist at P even in the absence of the mass M .

It is sometimes convenient to introduce the gravitational potential which determines the field through its gradient. For a spherical earth, it is defined as

$$\phi = -\frac{GM_E}{R}, \quad \mathcal{G} = -\text{grad } \phi$$

In general ϕ will satisfy Poisson's equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 4\pi\rho \quad (2)$$

ρ is the density of matter. The potential energy of a mass M , in the field is simply expressed in terms of ϕ ,

$$V = M\phi$$

Although one can introduce the gravitational field, it is an auxiliary concept in Newtonian theory for the field has no independent dynamical behavior as is true of the electromagnetic field (e.g., electromagnetic waves). At any time, the Newtonian gravitational field is determined by the configuration of masses at that instant and does not depend on previous history or state of motion. Thus if the sun were to vanish, the gravitational force on the earth would immediately be removed. This property may be thought of in terms of an infinite velocity of propagation for the gravitational field. Letting the velocity of light become infinite in Maxwell's equations eliminates all independent dynamical behavior for

the electromagnetic field. In that case there could be no radio or television. The special theory of relativity which is based on the velocity of light in vacuum being the maximum velocity for the transmission of energy, implies that Newton's theory requires modification.

Principle of Equivalence. The mass of a body may be measured either by weighing $W = Mg$ (*gravitational mass*) or by observing its motion under a known applied force using Newton's second law of motion $F = MA$ (*inertial mass*). The equality of these two differently defined masses has been measured by R. H. Dicke to an accuracy of $1 \cdot 10^{-11}$ improving an earlier measurement by Eötvös. It is this equality which distinguishes the gravitational force from all other forces in giving all bodies the same acceleration. The discussion of weightlessness pointed out that local effects of the gravitational field are eliminated for an observer in free fall precisely because all bodies fall at the same rate. It follows that the gravitational field *measured* by an observer will depend on his state of motion. In a sense *there is an equivalence between a gravitational field down and an acceleration up for the observer*. However, the equivalence is not complete, for real gravitational fields converge on their sources so that two particles released at the same time will drift closer together as they fall. On the other hand, acceleration fields have no effect on the separation of particles moving on parallel paths (Fig. 2). In a curved space, initially parallel geodesics—the "straight lines"—do not maintain a constant separation (e.g., great circles on a sphere). Thus, the gravitational field may have its explanation in the geometry of a curved space-time.

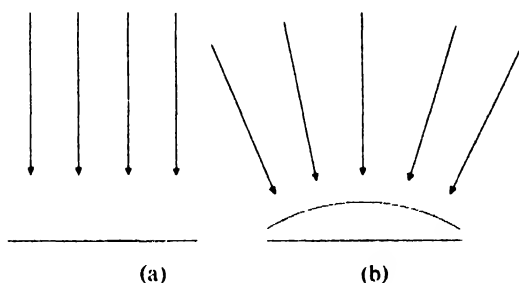


FIG. 2. The paths of particles released in

- (a) An acceleration field (the acceleration is up, the apparent force is down).
- (b) A gravitational field showing convergence toward the source.

Red Shift. According to the quantum theory, a photon of frequency ν has an energy $h\nu$ (h is Planck's constant), and by the relation $E = mc^2$, this quantum has a mass $m = h\nu/c^2$. To lift a mass m a height H requires expenditure of the energy mgH . Therefore, a photon emitted at the surface of the earth arrives at the height H with the energy

$$h\nu - (h\nu/c^2)gH = h\nu\left(1 - \frac{gH}{c^2}\right) = h\nu'$$

At the surface of the earth, the frequency shift amounts to

$$\frac{\Delta\nu}{\nu} = 1.1 \times 10^{-16}H \text{ (H in meters)}$$

This shift was measured by Pound and Rebka using the Mössbauer effect in good agreement with the prediction. As time standards are determined by frequency, it follows that if the same photon were emitted at the height H , it would be measured to have the frequency ν , not ν' . Therefore, an observer at H must conclude that his clock is running faster than the same clock would run on the surface of the earth in the ratio $\Delta T/T = -\Delta\nu/\nu$ (Fig. 3).

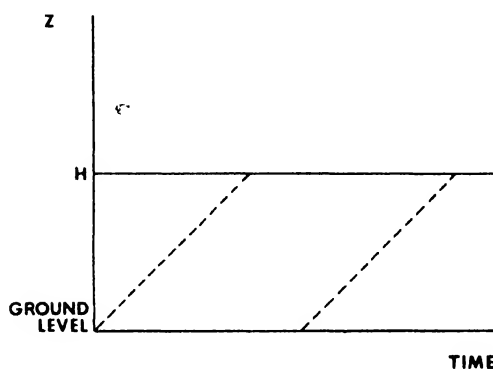


FIG. 3. Photons are emitted on the ground and are received at the height H . Between the two dotted lines representing the beginning and end of a pulse, the same number of oscillations, n , are received at H as are emitted at the ground level. Because of the red shift, the interval t' between oscillations at H is greater than the interval t between oscillations on the ground. Therefore, the time measured at H for the reception of the n oscillations is greater than the time required for their emission on the ground: $nt' > nt$. This result implies that clocks run faster at H than on the ground.

Einstein's theory of gravitation. Albert Einstein assumed that gravitation is a physical effect produced by the curvature of a four-dimensional space-time. The generalization of Newton's gravitational potential is the metric tensor $g_{\mu\nu}$ in terms of which the four-dimensional distance, and hence the geometry of space-time, is determined:

$$ds^2 = \sum_{\mu, \nu=1}^4 g_{\mu\nu} dx^\mu dx^\nu$$

The curvature of space-time is defined in terms of a four index tensor $R^\mu{}_{\nu\rho\sigma}$, the curvature tensor. The vanishing of the curvature tensor means that no real gravitational field is present. The field equations are ten linear combinations of the curvature components which are of the second order in the derivatives of the metric tensor and are a

generalization of Poisson's equation [Eq. (2)]. Symbolically these equations are written

$$G^{\mu\nu} = 8\pi\kappa T^{\mu\nu}$$

where $T^{\mu\nu}$ is a symmetric tensor which describes the distribution of matter and energy throughout space-time and $\kappa = G/c^2$. In a weak field static approximation, these equations contain Newton's theory of gravitation with the Newtonian gravitational potential. Given by $2\phi = 1 - g_{44}$.

The metric tensor outside a static spherically symmetric mass distribution is given by the Schwarzschild solution:

$$ds^2 = \left(1 - \frac{2\kappa m}{r}\right) dt^2 - \left(1 - \frac{2\kappa m}{r}\right)^{-1} dr^2 - r^2 d\theta^2 - r^2 \sin^2 \theta d\phi^2$$

This geometry exhibits the red shift described above and in addition shows two other effects:

(1) The bending of a ray of light passing near the sun's edge by

$$\delta\theta = 1.75''$$

(2) The precession of the perihelion of Mercury by

$$\delta\phi = 43''.03 \text{ century}$$

These results agree with the observations within the experimental errors which are fairly large—about 10 per cent.

Gravitational Waves. Einstein's field equations require that the gravitational field have a finite velocity of propagation—the same as that for light. Therefore, the gravitational field has independent dynamical degrees of freedom, which permit gravitational waves to exist in two modes. In passing through matter, one mode produces a compression along one axis and corresponding elongation along the perpendicular axis (Fig. 4); the other mode has the same effect along axes rotated by 45° . This character for the modes is caused by the tensor nature of the potentials $g_{\mu\nu}$, which limits the lowest order of gravitational waves to quadrupole radiation. A crude estimate of the energy radiated by the earth-sun system per year amounts to 10^{16} ergs (about 10^6 kWh). Radiating at this rate, the earth has lost about 10^{15} of its available mechanical energy since its formation 5×10^9 years ago.

Presumably there are stronger sources of gravitational waves available in the universe such as super novae, and the quasi-stellar sources. Experiments are in preparation by J. Weber of the University of Maryland to monitor gravitational waves received from the cosmos.

Quantum Theory. The gravitational interaction among the elementary particles is down by a factor of 10^{-40} from the electromagnetic or strong nuclear interactions. Therefore, one cannot expect to observe quantum effects at the level of today's experiments. Nonetheless, because the

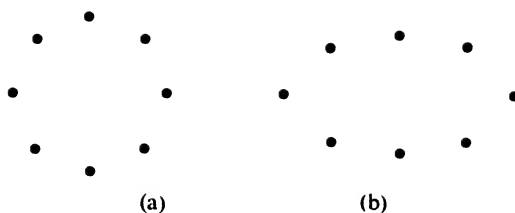


FIG. 4. (a). A circular arrangement of dust particles before a gravitational wave arrives.

(b). The same particles after a passage of a wave consisting of one mode. The second mode would produce the same effect, rotated at 45° .

gravitational field has two independent modes, the quantum field formalism requires that the gravitational field operators exist and satisfy appropriate commutation relations. Although there is no difficulty in principle, the gravitational field equations are complicated by being non-linear and by their covariance under arbitrary coordinate transformations. However, from its linear approximation, one expects that the quantized field will be a spin 2 boson field. Furthermore, because of its connection with the geometry of space-time, the gravitational field will be a link among all particles and all interactions.

Cosmology. One expects the gross structure of the universe to depend on global geometrical properties and the mean distribution of matter more than on the detailed kinetics of particle interactions. Among the various cosmological solutions of Einstein's equations are those models based on a homogeneous and isotropic matter distribution in which matter is streaming away from every point with a velocity proportional to its distance from that point. Such models agree with the observation made by Hubble which give a recessional velocity

$$v = Hr$$

with $H = 30$ km/sec/ 10^6 light-years and r measured in millions of light years. This value of H leads to an age for the universe of about 10^{10} years.

Recently some extremely luminous, relatively small, but very distant, objects have been observed. The most luminous of these quasi-stellar sources is 3C 273 at 1.6×10^9 light-years. This distance is so great that it is possible that a very early stage in the development of galaxies is being observed. The amount of energy being released is so great ($\sim 10^{46}$ ergs/sec) that a nuclear source for the energy is ruled out, and it is suggested that the energy comes from the gravitational contraction of $10^{10} M_\odot$ down to its characteristic radius of $2\kappa m = 2 \times 10^{10}$ km. This subject is under intensive research today.

JOSHUA N. GOLDBERG

References

Elementary

1. Gamow, George, "Gravity," *Sci. Am.*, **203**, No. 3, 94-100 (March 1961)
2. Gamow, George, and Cleveland, John M., "Physics, Foundations and Frontiers," Ch. 3, 7, and 20, Englewood Cliffs, N.J., Prentice Hall, Inc., 1960.
3. Einstein, A., "Relativity," New York, Henry Holt and Co., 1920.
4. Bonnor, William, "The Mystery of the Expanding Universe," New York, The Macmillan Co., 1964.

Technical

5. DeWitt, C., and DeWitt, B., Eds., "Relativity, Groups, and Topology," New York, Gordon and Breach, Science Publishers, 1964.
6. Bergmann, P. G., "The General Theory of Relativity," in "Encyclopedia of Physics," Vol. IV, Berlin, Springer-Verlag, 1962.

GYROSCOPE

The gyroscope consists of a flywheel or a sphere that is spinning (usually at high speed) about an axis. If this axis is free in space (such freedom may be provided by gimbals, by floating the spinning mass on a column of gas or fluid, or by suspension in a magnetic or an electrostatic field), the axis will remain parallel to its original position even though the gyroscope is mounted on a vehicle that translates and rotates in three dimensions. This property of the "free gyro" often referred to as "spatial memory" was used as an artificial horizon as early as 1744 by Serson. This permitted a ship's navigator to take readings with a sextant (measurement of the angular elevation of a star with respect to the horizon) when the horizon was obscured by darkness, fog mist, etc. This early instrument has led to the modern vertical gyro.

In 1852, Leon Foucault built one of the first precise gyroscopes. Utilizing a flywheel only a few inches in diameter supported in near-frictionless gimbals, this instrument was sensitive enough to detect the rotation of the earth. This free gyro maintained a fixed orientation in space as the earth turned. The relative motion was observed through a microscope.

In 1896, the gyroscope saw its first application for guidance when Obry used it in a self-propelled torpedo. An *unguided* torpedo, under the influence of winds, currents, and ocean waves would *not* follow a prescribed course. The gyroscope, on the other hand, was pointed at the target before launching the torpedo, and by means of linkages connected to the spin axis support, it actuated the rudder of the torpedo, steering it along a straight course.

When the gyroscope is *not* free, that is, when the spin axis is forced to turn in space, the gyroscope develops a torque about an axis that is perpendicular to the plane containing the spin axis and the axis about which turning takes place.

This property was utilized to find north by continually reorienting the spin axis until the gyroscope could *not* detect the *rotation* of the earth. This instrument, called the gyrocompass, was perfected in 1908 by Anschutz of Germany and by Elmer Sperry (1911) of the United States. This instrument used a wheel four to eight inches in diameter, and was suspended in such a way that it tended to remain vertical. Since the vertical turned with the earth, the gyroscope following it produced a torque. This torque, in turn, was used to provide self-turning of the spin axis in a direction that would reduce the torque. In the final equilibrium position, the instrument pointed north. The gyrocompass is still widely used today.

The gyroscopic torque is also used to stabilize ships. Due to the motion of waves, the *unstabilized* ship rolls considerably. It is impractical to shift huge masses fast enough to counteract the irregular motion. On the other hand, a gyroscope develops torque instantly and with much less effort. This led to the development of the gyroscopic ship stabilizer. Utilizing a wheel 10 to 20 feet in diameter, and a hydraulic turning mechanism, angular rates are applied to the spin axis support producing tremendous gyroscopic torques upon the ship. In moderately rough seas, the gyro stabilizer can eliminate 70 to 80 per cent of the roll motion. The control signal to the turning mechanism is provided by a small "vertical gyro" which is very much like Serson's artificial horizon gyro. This essentially is a guidance system (like Obry's) using a small guidance gyro to provide the vertical and a large gyro to provide the action.

If the motion of the gyro is free, and if a torque is applied to the structure containing the spinning mass, it will turn or precess about an axis that is perpendicular to the plane containing the torque axis and the spin axis. This is the converse of the gyroscopic torque phenomenon. By adjusting the amount of torque, the gyroscope provides a controlled rate. This property is utilized in an autopilot during a constant rate of turn. The principle is also used for platform stabilization where the platform carries instruments which must bear specific orientations with respect to the earth. By carefully regulating the torque, the gyroscope and platform rotate with the earth without being in contact with the earth.

Inertial navigation is accomplished by a gyroscopic system combining all the phenomena previously mentioned. Inertial navigation is required whenever visual or radiation methods cannot be used as the link with the earth. In its simplest form, this system consists of a free gyro platform that can be torqued. Prior to making the journey, the platform is leveled. As the ship, submarine, or missile circles the earth, accelerometers mounted on the platform continuously measure vehicular acceleration. This is integrated twice to determine the distance traveled. Dividing by the radius of the earth computes the angular position on earth (latitude and longitude). The gyroscope performs one of these integrations, and by appropriate scaling of the torque mechanism,

it divides the radius of the earth, tilting the platform through angles equal to the change in latitude and longitude. Thus, the system becomes an accurate analog of the motion of the vehicle on the earth.

In the more sophisticated inertial navigation systems, redundant measurements are made to minimize the growth of errors over long periods of time.

IRA COCHIN

References

Arnold, Ronald N., and Maunder, Leonard, "Gyrodynamics and its Engineering Applications," New York, Academic Press, 1961.

Cochin, Ira, "Analysis and Design of the Gyroscope for Inertial Guidance," New York, John Wiley & Sons, 1963.

Savet, Paul H., "The Gyroscope," New York, McGraw-Hill Book Co., 1961.

Cross-references: ROTATION-CIRCULAR MOTION, INERTIAL GUIDANCE; MASS AND INERTIA; MECHANICS.

H

HALL EFFECT AND RELATED PHENOMENA

If a current of particles bearing charges of a single sign and constrained to move in a given direction is subjected to a transverse magnetic field, a potential gradient will exist in a direction perpendicular to both the current and the magnetic field. This phenomenon is called the Hall effect, after E. H. Hall who discovered it in a metal in 1879.

Principles Involved. The Hall effect is a manifestation of the Lorentz force on a charged particle, which is expressed by the vector equation

$$\mathbf{F} = e[\mathbf{E} + (1/c)\mathbf{v} \times \mathbf{H}] \quad (1)$$

where \mathbf{F} is the force, e the charge, and \mathbf{v} the velocity of the particle, \mathbf{E} is the electric field, and \mathbf{H} is

the magnetic field intensity. A permeability of unity is assumed, and for the gaussian system of units, c is the speed of light. The physics involved in the Hall effect is illustrated by considering a confined stream of free particles, each having a charge e and an initial velocity v_x . A magnetic field in the z direction produces initially a deflection of charges along the y direction. This charge unbalance creates an electric field E_y and the process continues until the force on a moving charge due to the Hall field E_y counterbalances that due to the magnetic field so that further particles of the same velocity and charge* are no longer deflected. A pictorial representation of this is shown in Fig. 1. The magnitude of the Hall field follows at once from Eq. (1), namely

$$E_y = -(1/c)v_x H_z = J_x H_z / nec \text{ (gaussian}^\dagger \text{ units)} \quad (2)$$

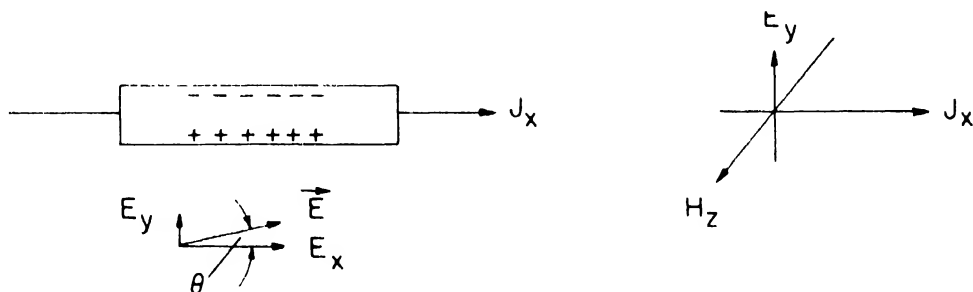


FIG. 1. Hall field due to action of magnetic field on positive charge carriers. For example shown, e , J_x and H_z are positive and, therefore, v_x , E_x , and E_y are also. The resultant electric field and the Hall angle are shown at lower left. In the case of electrons, the Hall field is reversed positive J_x .

* If the particles have a *distribution* of velocities, then it is only those particles having a certain "average" velocity, which are undeflected. This point is expanded later.

† In the gaussian system, mechanical quantities are in cgs units, electrical quantities in esu, and magnetic fields in gauss or oersteds. In the *practical* system, mechanical quantities are in cgs units, electrical quantities in volts and coulombs, and magnetic fields in gauss or oersteds. In all equations in this article, except for Eq. (1), conversion from the gaussian system to the practical system is effected by replacing c by unity and replacing H by $H/10^3$, where these quantities explicitly occur. The practical units for R_H are $\text{cm}^3/\text{coulomb}$; those for mobility are $\text{cm}^2/\text{volt-sec}$.

In obtaining the last equality, the electric current density J_x was expressed in terms of the density of charge carriers n by the product nev_x . For *electronic* (i.e., as opposed to *ionic*) conduction in solids, the magnitude of e is the electronic charge (1.6×10^{-19} coulomb, in practical units), and its sign is negative for transport by electrons and positive for transport by holes (deficit electrons). The *Hall coefficient* is defined by the ratio $E_y/J_x H_z$, namely,

$$R_H \equiv E_y/J_x H_z = -1/nec \text{ (gaussian units)} \quad (3)$$

Thus a very simple relation exists in the free-particle example between the Hall coefficient and the charge-carrier density. It is also seen that R_H

is negative for conduction by electrons, positive for conduction by holes. Now the electric current density J_x exists by virtue of an applied electric field E_x . With the Hall field present, the *resultant* electric field in a solid† lies at some angle θ to the x axis. This angle is called the Hall angle, namely

$$\theta \equiv \tan^{-1} E_y/E_x \quad (4)$$

Thus the Hall effect may be described as a rotation of the electric field. At zero Hall field, the equipotential lines are perpendicular to J , but when the Hall field appears, they are oblique, so that a Hall voltage exists across the specimen in a direction normal to the current. The rotation aspect is also brought out by considering the components of the *conductivity tensor*, which relate electric current densities and fields. For the boundary condition that $J_z = 0$, we may write

$$\begin{aligned} J_x &= \sigma_{xx}E_x + \sigma_{xy}E_y \\ J_z &= 0, H = H_z \quad (5) \\ J_y &= -\sigma_{xy}E_x + \sigma_{xx}E_y \end{aligned}$$

Equations (5) hold for media of sufficient symmetry* that $\sigma_{xx} = \sigma_{yy}$ and that $\sigma_{xz} = \sigma_{yz} = \sigma_{zx} = \sigma_{zy} = 0$. The latter insures that E_z vanish when J_z is zero. Since the boundary conditions for Hall effect require that J_y vanish, Eqs. (4) and (5) yield the following result for the off-diagonal elements σ_{xy} :

$$\sigma_{xy} = \sigma_{xx} \tan \theta \quad (6)$$

The inverse of the conductivity tensor is the *resistivity tensor*, which relates E to J . A general definition of the Hall effect involves relating it to the antisymmetric†† components of the resistivity tensor. This leads to the vector equation

$$E = R_H H \times J \text{ (gaussian units)} \quad (7)$$

The identity in the first part of Eq. (3) is recognized as a special case of Eq. (7).

Application to Real Solids. In a real solid, the idealized free particle treatment no longer applies, and it must be replaced by a theory that takes into account the distribution of velocities and the interactions of the charge carriers with impurities, defects, and lattice thermal vibrations of the solid—i.e., the *scattering*—as well as the *band structure* of the solid. The latter consideration relates to the fact that the charge carriers in the solid are not free but exist in a potential energy

field having the periodicity of the lattice. As a result of these constraints, only certain energy states, or *bands*, are allowed for the charge carriers. In addition, the relationship between the energy and velocity of the carriers is not the simple $\frac{1}{2}mv^2$ of the free electron, but is more complex. As an approximation, one frequently characterizes the charge carriers by an *effective mass*, m^* . Taking into account most of the complexities mentioned above, one still obtains an expression similar to Eq. (3), namely

$$R_H = r/nec \text{ (gaussian units)} \quad (8)$$

The *Hall coefficient factor*, r , actually depends on the nature of the scattering, the band structure, the magnetic field strength, and on the statistics characterizing the distribution of velocities of the carriers. Fortunately it depends weakly on these factors, and its value is usually within, say, 50 per cent of unity.

An important attribute of charge carriers is their mobility, i.e., their drift velocity per unit electric field. The conductivity mobility μ is related to the conductivity σ by

$$\mu = \sigma/nec \quad (9)$$

It is also customary to define a *Hall mobility* μ_H for conduction by a single type of charge carrier by the relation

$$\mu_H H/c = \tan \theta, \text{ or } \mu_H = R_H \sigma c \text{ (gaussian units)} \quad (10)$$

With the use of Eqs. (8) and (9), it can be seen that the ratio of Hall and conductivity mobilities is precisely the Hall coefficient factor. It follows that the Hall angle is proportional to μ and H .

When transport by two or more kinds of charge carriers occurs, Eq. (8) is not applicable, and one must return to the general relations of Eq. (5). For conduction by electrons and holes (of respective densities n , p), the *weak-field* Hall coefficient can be written in the form

$$R_H = -(r_e \mu_e^2 n - r_h \mu_h^2 p) / [e]c(\mu_e n + \mu_h p)^2 \quad (11)$$

where $|e|$, μ_e , etc., are positive. We note that the charge-carrier densities are now weighted by the mobilities μ_e , μ_h . There are also Hall coefficient factors for each carrier. For arbitrary H , the terms involve field-dependent factors.

Analysis of Hall-effect data is one of the most widely used techniques for studying conduction mechanisms in solids, especially semiconductors. For the single-carrier case, one readily obtains carrier concentrations and mobilities, and it is usually of interest to study these as functions of temperature. This can supply information on the predominant charge-carrier scattering mechanisms and on activation energies, i.e., the energies necessary to excite carriers from impurity levels into the conduction band. Where two or more carriers are present, the analysis is more complicated [cf. Eq. (11)], but much information can

† Although the Hall effect can be discussed for any constrained electron gas, we shall, for simplicity, talk about a solid conductor.

* Isotropic media, e.g., or cubic systems with coordinate axes along cube axes.

†† These components are defined by the condition that $\rho_{ik} = -\rho_{ki}$, where the ρ_{ik} [cf. Eq. (5)] are defined by $E_i = \sum_{k=1}^3 \rho_{ik} J_k$, $1 = x, 2 = y, 3 = z$.

be obtained from studies of the temperature and magnetic field dependencies.

Unlike, for example, the magnetoresistance, the Hall effect is a first-order phenomenon. At weak magnetic fields, it depends linearly on H , and it does not vanish in isotropic solids if all the carriers have essentially the same velocity or if the scattering is characterized by a relaxation time which is independent of the carrier energy. The Hall effect forms the basis of a number of devices used in isolating circuits, transducers, multipliers, converters, rectifiers, and gaussmeters (for measurement of magnetic fields).

Experimental Determination. A number of techniques are available, the most direct being to measure, by means of a potentiometer or other high-impedance device, the Hall voltage V_H across a parallelepiped in a direction normal to both H and J . The Hall coefficient follows from Equation (3) or (7):

$$R_H = 10^8 V_H t / IH \text{ (practical units)} \quad (12)$$

where t is the thickness of the specimen and I is the total current. The arrangement is shown in Fig. 2. Since V_H may be the order of microvolts, extreme care must be taken to avoid extraneous voltages. An example is the misalignment voltage, caused by probes 3 and 4 not being on an equipotential plane when $H = 0$. This may be eliminated by taking measurements for opposite directions of H and taking half the difference, inasmuch as $V_H(H)$ is odd in H . Alternatively, one can adjust the position of the Hall probes for a null reading with $H = 0$. Other techniques involve use of resistances with sliding contacts suitably attached to the specimen. Other spurious voltages can arise from temperature gradients and resulting thermoelectric or thermomagnetic emf's (see following section). Some of these can be eliminated by taking appropriate averages among measurements for reversed polarities of I and direction of H . It is usually desirable to maintain good thermal contact between all points of the specimen and a constant-temperature bath. Although it is possible to analyze *adiabatic* Hall

data (no heat flow to or from the specimen during measurement), *isothermal* data are preferred because of the simplicity of the equations. All of the relations in this article are for isothermal conditions. Errors in measurement can also result from a shorting of the Hall voltage, especially by the end contacts. With regard to the latter, the error is essentially negligible if the length-to-width ratio of the specimen is about 4 or more, and the Hall probes are near the center.

Related Effects. A widely-studied galvanomagnetic effect is *transverse magnetoresistance*, usually written $\Delta\rho/\rho_0$. This phenomenon can be illustrated by the pictorial scheme in Fig. 1. If a charge carrier is deflected by the Lorentz force so as to traverse a longer path, it will contribute less to the conductivity, and there will be a positive magnetoresistance. It was noted, however, that if all charge carriers have the same velocity, none will be deflected since the Hall field cancels the $\mathbf{v} \times \mathbf{H}$ force. This is the case in an isotropic metal, where the velocity of all the electrons is essentially the Fermi velocity. It is also true for electron scattering mechanisms described by a relaxation time which does not depend on energy. In these cases—assuming, of course, a single type of carrier—there is no magnetoresistance. If, however, there is a distribution of electron velocities—as in a semiconductor—then it is clear that only those electrons of a certain "average" velocity will be undeflected. The remaining carriers having velocities either larger or smaller than the "average" will be deflected and will traverse longer paths, thus increasing the resistance of the conductor. A similar situation obtains if more than one type of carrier is present. It is also apparent that any mechanism which shorts out the Hall voltage—e.g., special geometry, shorting contacts, inhomogeneities in the specimen—will increase the magnetoresistance.

For the reasons discussed and the fact that $\Delta\rho/\rho_0$ varies as H^2 in weak fields, magnetoresistance is a second-order effect. It tends to saturate in strong magnetic fields, unless there is a disturbance of the Hall field as mentioned above.

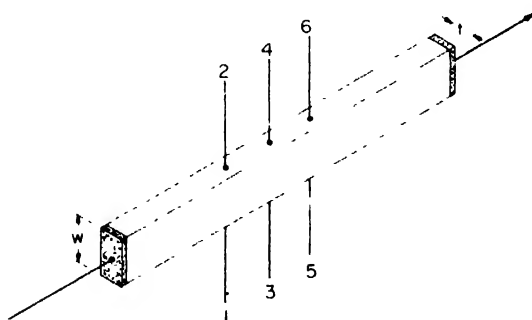


FIG. 2. Arrangement of contacts for measurement of Hall coefficient and related effects. Hall probes are 3 and 4. Probes 2 and 6 are for resistivity and magnetoresistance. Probes 1 and 5 allow a check of the uniformity of the specimen. For Hall effect and transverse magnetoresistance, the magnetic field is in the direction of the thickness t ; for longitudinal magnetoresistance, it is along l . To avoid disturbances due to contact shorting, all probes should be at a distance of at least $2w$ from the end contacts.

Magnetoresistance is even in H , and it is related to the *symmetric* components of the resistivity tensor, i.e., those for which $\rho_{ik} = \rho_{ki}$ [cf. footnote on p. 304]. It can be measured with the geometry shown in Fig. 2 by maintaining the current constant and determining the potential difference between probes 1 and 5 or 2 and 6 as a function of magnetic field. Magnetoresistance data can supply information on charge-carrier scattering and band structure. In the latter case, if anisotropy exists, it is useful to obtain data for different crystallographic directions. There is also a *longitudinal* magnetoresistance, measured when H is parallel to J . This effect vanishes in isotropic solids, and its presence indicates anisotropy in scattering or band structure, or inhomogeneities in the specimen.

By shorting the Hall field or by choosing a disk geometry so that such a field does not exist, one obtains a "magnetoresistance" (more strictly, a *magnetoconductivity*) which does not saturate. This is called the Corbino magnetoresistance or Corbino effect, after O. M. Corbino who studied circulating secondary currents in a "Corbino" disk carrying a primary radial electric current in a magnetic field.

There are a number of thermal effects in a magnetic field, which can produce transverse voltages or temperature gradients. These result from the velocity separation of charge carriers by the Lorentz force—the energetic ones going to one side, the slower ones going to the other. Temperature gradients are produced, and also electric fields. In the Righi-Leduc effect a longitudinal temperature gradient produces a transverse temperature gradient (thermal analog of the Hall effect); in the Nernst effect, it produces a transverse electric field. In the Ettinghausen effect, a longitudinal electric current produces a transverse temperature gradient. This latter effect, if large, can disturb the Hall field since the potential probes and leads are seldom made of the same material as the specimen. Therefore the Ettinghausen temperature gradient can produce a thermoelectric voltage which adds to the Hall voltage.

ALBERT C. BEER

References

- Beer, A. C., "Galvanomagnetic Effects in Semiconductors," (Supplement 4 to "Solid State Physics," F. Seitz and D. Turnbull, Eds.), New York, Academic Press, 1963.
- Dunlap, W. Crawford, Jr., "An Introduction to Semiconductors," New York, John Wiley & Sons, Inc., 1957.
- Fritzsche, Hellmut, "Galvanomagnetic and Thermomagnetic Effects," in "Methods of Experimental Physics," (L. Marton, Ed.), Vol. 6B, "Solid State Physics" (K. Lark-Horovitz and V. A. Johnson, Eds.), p. 145, New York, Academic Press, 1959.
- Lindberg, O., *Proc. Inst. Radio Engrs.*, **40**, 1414 (1952).
- Putley, E. H., "The Hall Effect and Related Phenomena," ("Semi-Conductor Monographs," C. A. Hogarth, Ed.), London, Butterworths, 1960.

Seitz, Frederick, "The Modern Theory of Solids," New York and London, McGraw-Hill Book Company, Inc., 1940.

Shockley, William, "Electrons and Holes in Semiconductors," Toronto, New York, and London, D. Van Nostrand Company, Inc., 1950.

Cross-references: CONDUCTIVITY; ELECTRICAL; ELECTRICITY; POTENTIAL; SEMICONDUCTORS; THERMOELECTRICITY.

HEALTH PHYSICS*

Health physics is the profession that is concerned solely with the protection of man from the damaging effects of ionizing radiation. It attempts to understand the action of radiation on man and his environment, to establish appropriate limits for exposure to radiation, and to devise appropriate methods for detection, measurement and control of radiation exposure. Although this profession is relatively new, man's awareness and concern for the harmful effects of ionizing radiation are not of recent origin. Perhaps the earliest record of damage to man from ionizing radiation dates back to about 1500 when the high incidence of lung diseases was recognized among the Schneeberg miners of Saxony and the Joachimsthal miners of Bohemia. In 1879 Herting and Hess performed the first autopsies on these miners and reported malignant growths in the lungs; however, the cause of these malignancies was not understood until after 1896 when Roentgen first announced his discovery of x-rays and Becquerel reported the discovery of radiation due to uranium. Even during the first year following the discovery of x- and γ -radiation, many things were learned about both the harmful and useful characteristics of this new source of energy. Grubbé, a manufacturer of Crookes tubes in Chicago, Illinois, was using his equipment to study the fluorescence of chemicals even before the public press on January 4, 1896, heralded Roentgen's discovery of x-rays. During January, 1896, he first noticed an erythema on the back of his hand and later the formation of a blister with skin desquamation and epilation. His hand was sufficiently painful that he sought medical aid on January 27, 1896. Realizing from first-hand experience the destructive power of x-rays, Grubbé on January 29, 1896, treated a patient for carcinoma of the breast with his Crookes tube. Not only was this treatment significant because it was one of the earliest - if not the first—therapeutic uses of ionizing radiation, but it is noteworthy that he acted as one of the first health physicists when he used lead as a shield to protect the rest of the body of the patient. Even Becquerel and Madame Curie learned from first-hand experience the need for radiation protection when they received skin

* Work sponsored by the U. S. Atomic Energy Commission under contract with the Union Carbide Corporation.

burns from the careless handling of radium (see RADIOACTIVITY).

S. Russ in 1915 made a comprehensive series of recommendations for radiation protection to the British Roentgen Society, and if these recommendations had been heeded, many of the early radiation fatalities might have been averted. It was not until 1928, when the International Commission on Radiological Protection was formed and published the first set of recommendations for radiation protection, that there began to be widespread interest and concern for this problem. In the following year the National Council on Radiation Protection was formed and it has set the standards in the United States.

Beginning at the time of the First World War and continuing until about 1930, there were many unfortunate exposures to radium. Some of these were the result of therapeutic injections of radium, the drinking of radium and radiothorium water, and occupational exposures of radium chemists. During this period, radium was considered to be a useful therapeutic agent, and as a result it was administered by physicians in the United States to hundreds of patients. In some cases it was taken as a general tonic, and in others it was given as a curative agent for hypertension, anemia, arthritis, and many other human ailments—even for insanity.

By far the most serious exposures were to young women engaged in the radium dial painting industry. Some of these women ingested relatively large quantities of radium as a consequence of tipping brushes with their lips as they applied radium paint on the dials of clocks and watches. The total number of radium dial painters and others who took radium by mouth or injection and, as a result, died with readily detectable symptoms of radiation damage is not known. The first recorded fatality due to radium-induced cancer, resulting from exposure in the radium dial industry, was in 1925, and since that time histories of over 50 such cases have been recorded and studied in the United States.

In 1942–43 there was begun at the University of Chicago a program to explore the possibility of assembling a critical mass of natural uranium in such a way that a "pile" or NUCLEAR REACTOR could be operated for the production of the new element plutonium to be used in atomic weapons. A. H. Compton, the director, and his associates debated the wisdom of proceeding with this project because they realized that in a single large reactor there would be produced ionizing radiation equivalent to that from thousands of tons of radium. Yet all the radium that had been available to man throughout the world only amounted to about two pounds, and these men were acutely aware of the extensive suffering and death that had resulted from its misuse. They decided to call together a rather unique group of scientists to evaluate these health problems, to develop new instruments, equipment, and techniques and to establish radiation standards for the protection of nuclear energy workers. The group assembled was concerned with the *health*

of the workers and consisted mostly of *physicists*; hence, they were called *health physicists*. E. O. Wollan was the leader and the other senior members were H. M. Parker, C. C. Gamertsfelder, K. Z. Morgan, R. R. Coveyou, J. C. Hart, L. A. Pardue, and O. G. Landsverk. Thus in 1942–43 health physics had its beginning at the University of Chicago. Although prior to this time many early pioneers such as S. Russ (England), L. S. Taylor (United States), G. Failla (United States), A. Mutscheller (Germany) and R. Sievert (Sweden)—to name only a few—had devoted considerable attention to the radiation protection problem, it was not until the advent of health physics that a professional group was organized with this as its sole objective.

As the nuclear energy programs expanded, large laboratories were established to carry on the program at Oak Ridge, Los Alamos, Hanford, Berkeley, Brookhaven, Savannah River, etc., and as the need for health physicists was recognized beyond nuclear energy programs and in private industry, hospitals, military organizations, state and federal agencies of public health and in colleges and universities, the profession of health physics grew and expanded very rapidly so that today (1965) it is estimated there are about 5000 health physicists in the world. In 1956 the Health Physics Society was organized which now has a membership of over 2500 in 43 countries. This Society publishes the journal, *Health Physics*. In 1959 the American Board of Health Physics was formed for the certification of persons whose technical competence and judgment qualify them to be responsible for handling major problems of radiation exposure and or contamination control.

There are three principal areas in which health physicists are employed—education and training, applied activities, and research—and these three areas will be discussed below.

In the early period, health physicists were scientists who had to develop their own competence during employment on the various atomic energy programs. In 1948 AEC Health Physics Fellowship programs were established, first at Oak Ridge National Laboratory and later at Vanderbilt University, University of Rochester, University of California, University of Washington, Harvard University, University of Kansas, University of Michigan, University of Puerto Rico, University of Tennessee and at Texas A & M. By 1964 a total of over 700 students had been awarded these Fellowships, and this program has become the principal source of senior health physicists, having produced approximately 300 at the masters level, 70 with Ph.D. degrees and over 350 with graduate training but without advanced degrees. Although most of the early health physicists began as physicists, there is today a need for health physicists with many different backgrounds, e.g., physics, biology, chemistry, mathematics, engineering, etc. In the AEC Health Physics Fellowship program the student meets all the usual Ph.D. requirements of courses, research and thesis in one of these major departments, and at the

same time, he meets additional requirements such as special courses and summer work at one of the National Laboratories where he is given practical experience in health physics. Likewise, some of the graduate programs of the U.S. Public Health Service have been, in more recent years, providing training in health physics. In spite of the various opportunities for education and training in health physics, the supply has not kept pace with the demand because of the rapid growth of this new profession.

In addition to the above-mentioned graduate programs in health physics, there are education and training programs at all operating levels. These programs are important because the success of health physics can be measured, to a considerable degree, in terms of how well plant managers, supervisors, scientists, engineers, technicians and operations personnel are made to realize their responsibility for protecting themselves and their associates from radiation damage. They must be ever aware that to some degree all radiation exposure is harmful and no unnecessary exposure may be permitted unless it can be balanced by benefits of equal value. At the same time they must be made to respect ionizing radiation—not fear it.

The duties of the health physicist in applied operations are very diverse and differ considerably from place to place, depending upon the size and nature of the operation. For example, the health physicist in a reactor operation would have duties quite different from those of the health physicist associated with an accelerator program or the health physicist connected with a state public health organization charged with the survey of medical x-ray equipment. A few typical applied health physics activities may be summarized as follows:

- (1) Aid in the selection of suitable locations for buildings in which radioactive materials are to be produced or used, and conduct pre-operation background surveys.

- (2) Offer advice in the design of laboratories, hoods, remote control equipment, radiation shields, etc.

- (3) Provide personnel monitoring meters for radiation dosimetry to all persons subject to radiation exposure and read these meters frequently; make thyroid counts, breath measurements, urine and feces analyses; check body with scanners and total body counters and conduct other tests to aid in estimating how much (if any) radioactive material is fixed in the body; and maintain accurate records of the accumulated dose from each type of ionizing radiation received by each individual for his protection and for the protection of the employer.

- (4) Make frequent surveys of all accessible reactor areas, radioactive sources, x-ray equipment, high voltage accelerators, chemistry and physics laboratories, metallurgical shops, and other working areas where radiation exposure is possible.

- (5) Advise scientists, supervisors and research directors of all radiation exposure hazards, of

permissible working time in a given area, and of radiation protection measures and techniques (e.g., protective clothing, shields, remote control equipment) and aid supervision in the solution of new radiation problems as they develop.

- (6) Make frequent surveys of all radioactive waste discharged beyond the area of immediate control, maintain accurate records of the level of this radioactivity in the air, water, soil, vegetation, milk, etc., and advise management of remedial measures as they are needed.

- (7) Aid in all emergency operations where there are associated radiation hazards.

- (8) Purchase and maintain in working order and in proper calibration suitable health physics survey and monitoring instruments which are used as aids in the protection of personnel from radiation damage.

- (9) Prepare operations manuals on "Rules and Procedures Governing Radiation Exposure."

- (10) Assist in radiation protection problems related to civil defense, weapons fallout, the use of nuclear devices for excavation of canals and harbors, space radiation, etc.

Health physics research ranges from the applied and engineering programs to very basic studies. It is a working together of scientists of many disciplines—physicists, chemists, biologists, engineers, geologists, mathematicians—all studying the effects of ionizing radiation on man and on his environment. In these studies, they are working at all levels—nuclear, atomic, molecular, plasma, gas, solid, liquid, cell, animal, and the ecosystem. In this research program, radiation ecologists are studying the effects of low levels of radiation exposure on the environment. Some essential organisms in the environment are known to concentrate radioactive waste by a factor of 10,000 or more, and the health physicist must determine the importance of the indirect damage of ionizing radiation to man's environment as well as its direct effects. Internal dose studies are under way by researchers in health physics, studies which have led to the publication of the official handbooks on maximum permissible concentration of the various radionuclides in food, water, and air. These handbooks are issued by the National Council on Radiation Protection and the International Commission on Radiological Protection. They are under constant revision by the health physicist as more reliable and detailed information becomes available. Biologists in health physics are studying the uptake distribution and elimination of radionuclides which are taken into animals and man by the several modes of intake—ingestion, inhalation and skin penetration. Engineers in health physics are exploring and demonstrating new methods for the disposal of radioactive waste in deep wells thousands of feet below the earth's surface and in salt mines. They are studying the seepage rates of radionuclides into various soil formations, its slow dissipation from packages of radioactive waste deposited on the ocean floor, and the dilution of airborne radioactive waste as it is discharged from stacks under varying meteorological

conditions. The physicists in health physics are making basic studies of the various energy exchanges that take place in matter when it is exposed to ionizing radiation. This information aids in the development of better radiation detection systems and leads to an understanding of the true meaning and consequence of radiation exposure. When high energy radiation (in the MeV or keV region) strikes living matter, there are innumerable, complex energy exchanges that take place as the entropy of the system increases. An understanding of the many low energy transitions is basic to a proper interpretation of the effects of ionizing radiation on man. Health physicists are working toward the ultimate goal of developing a coherent theory of radiation damage. Only when such a theory is available can they have complete understanding and confidence in the many extrapolations to man of the effects of ionizing radiation on animals. Such information is essential in developing reliable radiation protection standards and measures that are enforced by the applied health physicist.

Health physics continues to grow as a most interesting and challenging profession for scientists of many backgrounds. The success of these programs is attested by the fact that ionizing radiation with unparalleled potential for radiation hazards has expanded in its use and applications into almost every area of human endeavor and yet the nuclear energy industry has become one of the safest of all industries.

KARL Z. MORGAN

Cross-references: MEDICAL PHYSICS, NUCLEAR INSTRUMENTS, NUCLEAR RADIATION, NUCLEAR REACTORS, NUCLEONICS, RADIOACTIVITY, REACTOR SHIELDING.

HEARING

The role of the sense organ of hearing is to code acoustic disturbances into neural signals suitable for transmission to the brain. The study of this process necessarily involves anatomy and physiology of the ear, the nature of auditory pathways and central nervous system activity in hearing, properties of acoustic signals that elicit auditory responses, and observed phenomena of auditory behavior. These aspects serve to define and delineate areas for investigations of hearing and are the topics of discussion for this article.

In this approach to hearing, questions are asked about the structure of the system, how the system functions, and the relationships between inputs and outputs of the system. These three kinds of data—morphological, physiological, and psychological—need to be compared and correlated for a full understanding of hearing. It is important, however, that these three frames of reference be kept separate and not be confused. Although physiological functions may correspond in a general way to anatomical sequences, several physiological functions may occur in the same anatomical structure or a single function may require several anatomical units. In a similar manner, psychological functions cannot usually

be identified with specific physiological functions, and it is recognized that the central nervous system, as well as the auditory system, is involved in any auditory response. The correlations between and knowledge about structure and functions are best developed for peripheral, rather than central, parts of the auditory system because the ear is more accessible for examination and study than are the more central parts of the auditory system.

Traditional theories of hearing have been largely concerned with pitch perception or loudness of pure tones. Recently, there has been an increased awareness that any comprehensive theory of hearing needs to encompass various experimental phenomena of hearing. In this regard, increased attention has been given to auditory processing of complex signals, to the examination of binaural inputs to the system, and to the study of pathological hearing conditions.

When the structure of the ear is examined, it is convenient to consider the external, middle and inner ear separately. From a functional point of view, the ear may be divided into an outer and inner part. The outer is concerned with the transformation of acoustic energy into mechanical energy and the inner with the transduction of mechanical energy into neural impulses. The auricle and external auditory meatus constitute the external ear. The meatus is an irregularly shaped tube approximately 27 mm long with a diameter of about 7 mm, terminated by the tympanic membrane. The ear canal is an acoustic resonator, and frequencies in the range of 3000 to 4000 cps are increased in pressure at the ear drum as compared to the pressure at the entrance to the canal. The ear drum is in a protected position at the end of the canal, and humidity and temperature conditions at the drum are relatively independent of those external to the ear.

The middle ear is an irregular, air-filled space in the petrous portion of the temporal bone. The three ossicles of the middle ear—the malleus, the incus, and the stapes—provide mechanical linkage between the tympanic membrane and the fenestra vestibuli, an opening in the vestibule of the inner ear commonly known as the oval window. The handle of the malleus attaches to the tympanic membrane, and the footplate of the stapes attaches to the oval window. Two important functions are provided by the middle ear. The first is to amplify and deliver sound vibrations from the drum to the inner ear, and the second is that of protecting the inner ear from very loud sounds. The amplification of sound waves is accomplished by apparent lever action of the ossicles that produces a greater force at the oval window than the force at the drum and because of the gain in force that results from the relationship between the larger drum area to the smaller stapedial footplate area. The area of the drum is approximately 25 times that of the oval window. The amplification gain of these two factors is approximately 25 db. The effectiveness of the middle ear action in increasing hearing sensitivity is evidenced in middle ear pathologies where the ossicular chain is disrupted. A hearing

loss of 25 db or more occurs. The second function of the middle ear, that of protecting the inner ear from loud sounds, is accomplished by reflex action of the middle ear musculature, the tensor tympani, and the stapedius. The action of the muscles is to retract the ear drum, draw the stapes away from the oval window, and change ossicle vibrations in such a way as to decrease the transmitted pressure. Latency of muscle contraction and possible muscle fatigue limit protection of the inner ear by these mechanisms. Middle ear air pressure is equalized by virtue of the Eustachian tube which connects the middle ear and the nasopharynx. The pressure equalization is necessary for normal ear drum movement.

The inner ear is a system of cavities in the dense petrous portion of the temporal bone. One of the cavities is the cochlea, a bony labyrinth that is approximately 35 mm in length coiled around a central core for two and three-quarters turns. The spiral-shaped cochlea is divided into three ducts, two bony and one membranous. The upper bony duct, the scala vestibuli and the lower bony duct, the scala tympani, are separated from each other by a membranous labyrinth, the cochlear duct. The cochlear duct is bound on top by Reissner's membrane and is bound below by the basilar membrane. The cochlear duct is filled with a viscous fluid called endolymph, and the duct is surrounded by a fluid called perilymph that has about twice the viscosity of water. The scala vestibuli and the scala tympani join at the apical end of the cochlea at a passage called the helicotrema. The scala tympani terminates at the basal end at the round window, a membrane-covered opening into the middle ear. The scala vestibuli is continuous with the vestibule; the oval window opens into the vestibule. Vibrations at the footplate of the stapes are transmitted into the fluid adjacent to the oval window. Vibration of the stapes and resultant disturbances in cochlear fluids results in movement of the basilar membrane. The cochlear duct contains the sensory receptors, specifically the organ of Corti, which lies upon the basilar membrane. There are about 25 000 hair cells; one end of each rests on the basilar membrane. The other ends of the hair cells are the cilia, very fine hairline processes, which make contact with the tectorial membrane, a membrane that overlaps the organ of Corti and that functionally behaves as if it were hinged at the cochlear wall. There are three rows of outer, and one row of inner, hair cells along most of the length of the basilar membrane. When vibrations are introduced into the inner ear and cause displacement of the basilar membrane, a shearing of the action of the cilia occurs that results in neural activity. It is assumed that amplification occurs in the inner ear in that small pressures on the basilar membrane result in a shearing force of considerably greater magnitude that distorts the hair cells. The result is increased sensitivity of the hearing system. Physical properties of the cochlea are such that different frequencies tend to localize at different points along the basilar membrane. The basilar membrane is narrowest and stiffest at the basal end,

and most lax and widest at the apical end of the cochlea. High-frequency sounds result in the greatest disturbances near the basal end, and low-frequency sounds tend to localize near the apical end. When the role of the cochlea in pitch and loudness analyses is considered it is now realized that more is involved in pitch perception than the place of localization on the basilar membrane, although the particular neural fibers involved are probably relevant. Loudness is probably related to the total number of neural impulses per unit time.

The auditory pathways provide for the neural impulses from the ear to be transmitted to the cerebral centers of the auditory cortex. Processing of the neural signals probably occurs at synaptic connections as well as in the cortex. The cell bodies of the receptor neurons are located in the spiral ganglion. Neurons of the auditory nerve make synaptic connections with the hair cells of the cochlea. Nerve fibers typically innervate many hair cells, and more than one nerve fiber may make a connection with the same hair cell. There is recent evidence to indicate that there are also descending neural pathways as well as ascending ones. The central nervous system may thus be involved in auditory processing at the cochlea. Spiral ganglion axons make synaptic connections with cells of the central nervous system at the cochlear nucleus. At this point, there is interconnection between the pathways for the two ears. Other synaptic stations between this point and the auditory cortex include the inferior colliculus and the medial geniculate body. Evidence from pathological auditory systems is of particular interest with respect to the auditory pathways. An impaired cochlea, for example, may result in a better than normal response to small amplitude changes in a sound. A lesion of the VIIIth Nerve is frequently manifested by a rapid decrease in the ability to respond under sustained stimulation. The ability to process speech is markedly affected when there is an involvement of the lower central nervous system. Cortical involvement does not affect usual speech or pure tone inputs.

Sound involves a disturbance in the air that is a forward and backward, rarefaction and compression, movement of air particles. The unit of force usually used in acoustics is the dyne. Sound pressure is frequently expressed in dynes per square centimeter. Intensities of sounds are usually measured on a decibel scale, a logarithmic ratio scale. The tremendous loudness range of the ear is exemplified by the fact that the most intense sound that can be tolerated is a million million times greater in intensity than a sound that is just audible. This is a range of approximately 120 db. The frequency range of hearing is frequently given as 16 to 20 000 cps. The ear is most sensitive in the middle frequency range of 1000 to 6000 cps. In terms of discrimination of frequency and intensity, it is possible for about 1400 pitches and 280 intensity levels to be distinguished.

The truly phenomenal aspects of hearing can be observed in such behavior as localization of sounds, speech perception and particularly the

understanding of one voice in the noisy environment of many, and the recognition of acoustic events that only last a few milliseconds. It is these and other behavioral phenomena that need to be accounted for in theories of hearing.

ROBERT W. PETERS

References

Books

- Bekesy, Georg von, "Experiments in Hearing," New York, McGraw-Hill Book Co., 1960.
 Jerger, James, "Modern Developments in Audiology," New York, Academic Press, 1963.
 Rasmussen, Grant, Ed., and Windle, William F., "Neural Mechanisms of the Auditory and Vestibular Systems," Springfield, Charles C. Thomas, 1960.
 Stevens, S. S., "Handbook of Experimental Psychology," Chs. 27 and 28, New York, John Wiley & Sons, Inc., 1951.
 Stevens, S. S., and Davis, Hallowell, "Hearing: Its Psychology and Physiology," New York, John Wiley & Sons, 1947.
 Van Bergeijk, Willem, Pierce, John R., and David, Edward E., Jr., "Waves and the Ear," New York, Doubleday and Co., 1960.

Periodicals

- "The Bekesy Commemorative Issue," *The Journal of the Acoustical Society of America*, 34, 9 (September 1962) Part II

Cross-references: ACOUSTICS; ARCHITECTURAL ACOUSTICS; MUSICAL SOUND; NOISE, ACOUSTICAL; PHYSICAL ACOUSTICS; REPRODUCTION OF SOUND.

HEAT

Heat vs Temperature. Heat is the imponderable but not intangible agency whose addition to or removal from a physical system is the cause of thermal changes of various types. These include rise and fall of temperature, changes in length and volume, changes of physical states such as melting, evaporation and the like.

During the eighteenth century heat was assumed to be a subtle fluid called *caloric*, filling the interstices between the ultimate particles of matter and, under conditions of isolation from the surroundings, known to satisfy a conservation law. The production of heat by friction as well as its disappearance during the performance of external mechanical work established its essential physical nature as another form of *energy* and led to the overthrow of the caloric theory. Nevertheless, we still speak of the *flow* of heat as though it were a fluid and have retained the methods of measuring the *quantity of heat* originally devised by the upholders of the caloric view.

Our direct knowledge of heat is provided by the sensation of hotness and coldness when we come in contact with various physical bodies. It is possible to arrange a set of bodies in a sequence such that A feels hotter than B, B hotter

than C, etc. We say that A has a higher *temperature* than B, B a higher one than C, and so on. Of course our sensations are qualitative and are considerably influenced by the thermal conductivity of the body we touch. Thus, on a frosty morning, the head of an ax being metal feels considerably colder than the wooden handle though the two are presumably at the same temperature. To obtain a continuous and reproducible physical scale of temperature, various types of thermometers have been devised of which the mercury-in-glass or colored-alcohol-in-glass are familiar examples. The two temperature scales in common use are the Fahrenheit scale and the Celsius scale. The first assigns values of 32° and 212° to the normal freezing and boiling points of pure water, respectively, and divides this interval into 180 equal sub-intervals or degrees. The Celsius, formerly called the Centigrade, scale assigns the respective values of 0° and 100° to the above fixed points; the standard interval is then divided into 100 equal degrees.

Temperature changes being produced by the addition or subtraction of heat from a body, temperature itself may be regarded as a measure of the concentration or *intensity* of heat. In general, the more heat we add to a given body the more its temperature rises.

Measurement of Heat. Since heat is imponderable and not directly observable, it is necessary to measure the size of a given quantity of heat by its effect on another body. If this effect is the production of a rise in temperature from some initial temperature, t_1 , to a final temperature, t_2 , then the rise ($t_2 - t_1$) is found to vary inversely with the mass of the test body. It is thus natural therefore, following the calorists, to regard the quantity of heat, say Q , as determined by the product of m and ($t_2 - t_1$). Thus we say

$$Q \text{ is proportional to } m \cdot (t_2 - t_1)$$

To make this statement into an equation we write

$$Q = \text{constant} \cdot m \cdot (t_2 - t_1) \quad (1)$$

where the constant of proportionality depends on the substance, being large for some materials and small for others. This constant for water, for example, is about 33 times as great as for lead; water is said therefore to have a greater *heat capacity* than lead. Notice that the constant in Eq. (1) actually gives the numerical value of Q which is required to warm a unit mass of the substance through a temperature interval of exactly 1°. This constant is accordingly called the *specific heat capacity* (usually abbreviated to *specific heat*) and is indicated by c . Since it is found that the value of the specific heat, particularly for gases, but in principle for all materials, depends on the conditions under which the heat is absorbed, this must be indicated. We thus have c_p and c_v , for example, for the two important cases of absorption at constant pressure and constant volume respectively. Since the former characterizes the common laboratory case of

working under atmospheric pressure, we accordingly rewrite Eq. (1) as

$$Q_p = c_p m(t - t_1) \quad (2)$$

Q_p now measures the heat absorbed under constant pressure, and c_p is the constant pressure specific heat. Since the right side of Eq. (2) contains *three* quantities, a mere choice of a mass unit and a degree unit is insufficient to establish a unit of heat. It is necessary to select some substance as a standard reference body and assign an arbitrary value of, say c_p equal to unity for it. Water is the universal choice for this standard body due not only to its cheapness and ease of purification but also to its large heat capacity.

With the selection of water as the standard with $c_p = 1$, the left side of Eq. (2) clearly becomes of unit value when m and $(t - t_1)$ are each of unit value. In the English system we accordingly have the *British thermal unit* (or Btu) as the heat required to warm 1 pound of pure water through an interval of 1° F. In the metric system, the corresponding unit is the *calorie*, the heat required to warm 1 gram of water 1°C. A large or *kilocalorie* corresponding to 1000 ordinary calories is also frequently used in scientific work.

Specific Heats. Use of Eq. (2) reveals that the values of c_p obtained experimentally depend on the temperature interval used, indicating a dependence of c_p on temperature. Thus if c_p for water were actually unity throughout the 0 to 100°C range, a mass of water at 100°C mixed with an equal mass at 0°C would give a final mixture at exactly 50°C. The actual value is near 50.05°; this difference although small, indicates the need to specify the calorie at some particular temperature. For this purpose, we suppose a system of mass m is warmed from t to $t + \Delta t$ by the addition at constant pressure of an increment of heat ΔQ_p . Then Eq. (2) becomes

$$\Delta Q_p = m \bar{c}_p \Delta t \quad (3)$$

where now \bar{c}_p is an average value of c_p over this interval. Then we define the *instantaneous* heat capacity, c_p at t by the following relation:

$$c_p = \frac{1}{m} \lim_{\Delta t \rightarrow 0} \frac{\Delta Q_p}{\Delta t} = \frac{1}{m} \frac{dQ_p}{dt}$$

i.e., the heat absorbed per unit mass per degree as the interval becomes smaller and smaller without limit. This leads to the differential form of Eq. (3)

$$dQ_p = m c_p dt \quad (4)$$

where dQ_p is the differential heat absorption which produces a differential temperature rise dt in a body of mass m and specific heat c_p .

The standard or 15 calorie is now defined as the rate of absorption of heat per gram per degree at 15°C and in practice is essentially the same as the average calorie over the 1° interval from 14.5 to 15.5°C.

If a mass m of water is warmed from t_1 to t , the integral of Eq. (4) gives for the total heat absorbed in 15 calories

$$\begin{aligned} Q_p &= \int_{t_1}^t dQ_p = m \int_{t_1}^t c_p dt = \\ &= m \left[\int_{0^\circ}^t c_p dt - \int_{0^\circ}^{t_1} c_p dt \right] \quad (5) \end{aligned}$$

where the integral of c_p over the range t_1 to t has been written as the difference of two integrals from a common lower limit of 0°C. If, therefore, we evaluate an integral of the type $\int_0^t c_p dt$ with t varying in 1° steps and arrange these in a table, the right side of Eq. (5) may be evaluated by merely subtracting appropriate entries.

In Fig. 1, the value of c_p in 15° calories per gram per degree is plotted graphically from 0 to 100°C and the integrals on the right of Eq. (5) are represented by appropriate areas under the c_p curve. Thus the integral from 0° to t is hatched with lines sloping up to the right, while that from 0° to t_1 has the lines sloping up to the left. The value of Q_p is then the singly hatched area.

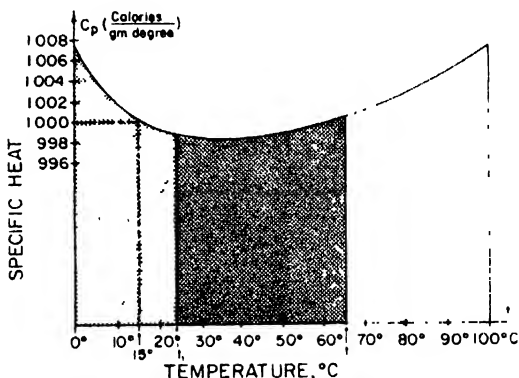


FIG. 1. Specific heat of water vs temperature.

With heat quantities measured in 15° calories, from the observed rise or fall of temperature in known masses of water, the specific heats of various substances, the heats absorbed on melting solids to liquids (heats of fusion), the heats absorbed on passage from the liquid to the vapor state (heats of vaporization), the heats evolved on combination of various substances, and the heats absorbed or evolved in chemical changes are at once determinable (see CALORIMETRY). For the present purpose, Table 1 gives the values of the constant pressure heat capacities of a few typical substances, variations with temperature being disregarded. Notice that c_p although expressed in terms of calories per gram per degree is in fact independent of the system of units since water is the reference body in all systems. Thus the specific heat of water in the English system would be 1 Btu per pound per degree Fahrenheit.

The Mechanical Nature of Heat. The conservation of heat *per se* is observed only for systems involving the performance of no mechanical or electrical work. Count Rumford (ca. 1800) was the first to establish this fact in his famous cannon-boring experiments carried out in the

TABLE I

Substance	State	$c_p(\text{cal/g deg})$
Water	Vapor	0.48
Water	Liquid	1.00
Water	Solid	0.50
Ethyl alcohol	Liquid	.54
Hydrogen	Gas	3.44
Air	Gas	.24
Aluminum	Solid	.22
Iron	Solid	.11
Lead	Solid	.03

arsenal of the Dutchy of Bavaria in Munich. He observed that when his drills became dull, heat was produced in great quantities limited only by the amount of work done against friction. He concluded that the large scale mechanical energy used in overcoming friction could only be converted into the motions of the ultimate particles of matter, a motion not directly observable but detected by our senses as heat. His results were confirmed and extended by the later work of Joule and Helmholtz, in particular, and also provided a more reliable value for the so-called *mechanical equivalent of heat*. This is taken as the amount of mechanical (or electrical) energy which when converted into heat is equivalent to exactly 1 calorie. The presently accepted value for this important constant is 4.185 joules per 15° calorie. Here the joule is the work performed when power is expended at the rate of 1 watt for 1 second. Thus an ordinary 100-watt lamp bulb converts 100 joules of electrical energy to thermal each second; this amounts to 100/4.185 or about 24 calories.

As a result of experiments such as these and a host of others, we are forced to recognize that heat is merely another form of the universal quantity *energy*. Its transformation always occurs at the rate of 4.185 joules per calorie whether heat goes into external work or work is dissipated through friction into heat.

For details of the mechanical interpretations of heat in the case of ideal gases and simple solids see KINETIC THEORY.

FRANZO H. CRAWFORD

Reference

Crawford, Franz H., "Heat, Thermodynamics and Statistical Physics," New York, Harcourt, Brace and World, 1963.

Cross-references: CALORIMETRY, HEAT CAPACITY, KINETIC THEORY, TEMPERATURE AND THERMOMETRY.

HEAT CAPACITY

The heat capacity of any thermodynamic system is

$$C = \frac{\delta q}{dT} \quad (1)$$

where δq is the quantity of heat* required to produce the temperature increment dT . If the system consists of a single substance with definite chemical composition and physical state, C is proportional to the total quantity of matter in the system. The heat capacity of 1 gram is the specific heat, and the heat capacity of one mole is sometimes called the molar heat capacity. Commonly used units of heat capacity include cal mole⁻¹ deg⁻¹, cal g⁻¹ deg⁻¹, and joule mole⁻¹ deg⁻¹.

The heat capacity of a substance depends on the variables that determine its thermodynamic state—e.g., temperature, pressure, electric and magnetic field—and also on the constraints imposed during the absorption of heat. Here we limit consideration to a substance subject to a hydrostatic pressure P and to no other forces. In this case, the thermodynamic state is determined by two variables, usually taken to be temperature and either pressure or molar volume V . The usual constraints for which C is of interest are constant volume and constant pressure, and the symbols C_V and C_P are used for the heat capacities measured under these conditions.

The difference between C_P and C_V is related to other thermodynamic properties by the first law of thermodynamics, which, for processes of interest here, may be written

$$\delta q = dU + P dV = dH - V dP \quad (2)$$

where U is the energy and $H = U + PV$ is the enthalpy. (We write this and all subsequent expressions involving thermodynamic properties for one mole.) From Eqs. (1) and (2), we have

$$C_V = \left(\frac{\partial U}{\partial T} \right)_V \quad (3)$$

and

$$C_P = \left(\frac{\partial U}{\partial T} \right)_P + P \left(\frac{\partial V}{\partial T} \right)_P = \left(\frac{\partial H}{\partial T} \right)_P \quad (4)$$

Subtraction of Eq. (3) from Eq. (4) and substitution of the mathematical relation

$$\left(\frac{\partial U}{\partial T} \right)_P = \left(\frac{\partial U}{\partial T} \right)_V + \left(\frac{\partial U}{\partial V} \right)_T \left(\frac{\partial V}{\partial T} \right)_P \quad (5)$$

gives

$$C_P - C_V = \left[\left(\frac{\partial U}{\partial V} \right)_T + P \right] \left(\frac{\partial V}{\partial T} \right)_P \quad (6)$$

The quantity of heat required to produce a certain temperature increase dT is greater at constant pressure than at constant volume by the sum of the work of expansion in the constant pressure process, $P(\partial V/\partial T)_P dT$, and the increase in internal energy accompanying the expansion, $(\partial U/\partial V)_T(\partial V/\partial T)_P dT$.

An expression for $C_P - C_V$ that is more useful than Eq. (6) is

$$C_P - C_V = TV\alpha^2/\kappa \quad (7)$$

* The symbol δ indicates a differential that is not exact, i.e., not determined by the initial and final thermodynamic states.

which can be obtained from Eq. (6) by introducing a relation based on the second law of thermodynamics,

$$\left(\frac{\partial U}{\partial V}\right)_T + P = T\left(\frac{\partial P}{\partial T}\right)_V$$

the mathematical relation

$$\left(\frac{\partial P}{\partial T}\right)_V = \left(\frac{\partial V}{\partial T}\right)_P / \left(\frac{\partial P}{\partial P}\right)_T$$

and the definition of the coefficients of thermal expansion α and isothermal compressibility κ ,

$$\alpha = \frac{1}{V} \left(\frac{\partial V}{\partial T}\right)_P \quad (8)$$

and

$$\kappa = -\frac{1}{V} \left(\frac{\partial V}{\partial P}\right)_T \quad (9)$$

The subject of the following sections is the relationship of the heat capacities of ideal gases and solids to the energies associated with different degrees of freedom of the constituent particles. The heat capacity of a liquid is more complicated and is not considered here.

Ideal Gases. An ideal gas obeys the equation of state, $PV = RT = NkT$. The gas constant R is 8.314 joule mole⁻¹ deg⁻¹, N is Avogadro's number, and $k = R/N$ is Boltzmann's constant. By Eqs. (7), (8) and (9), $C_P = C_V + R$.

In the application of statistical mechanics to the calculation of C_P or C_V it is convenient to consider first the predictions based on classical statistical principles. The pertinent result of classical statistical mechanics is the principle of equipartition of energy: any term in the energy of a particle proportional to the square of either a coordinate or a momentum contributes an average of $\frac{1}{2}kT$ to U . Since the molecules of an ideal gas have three translational degrees of freedom, each with an associated kinetic energy proportional to the square of a momentum, the expected translational contributions to the energy and heat capacity are $U = \frac{3}{2}NkT = \frac{3}{2}RT$ and, by Eq. (3), $C_V = \frac{3}{2}R$. For monatomic gases this is the only contribution to C_V and this value is in excellent agreement with experiment.

Diatomic molecules can rotate about each of two independent axes perpendicular to the internuclear axis, and the two atoms can vibrate with respect to each other along the internuclear axis. Each of the two rotational degrees of freedom has a kinetic energy proportional to the square of a momentum, and the vibrational degree of freedom has both a kinetic energy proportional to the square of a momentum and a potential energy proportional to the square of a coordinate (the vibration is approximately harmonic). In classical statistical mechanics, C_V is therefore expected to be $\frac{7}{2}R$, but for most diatomic gases at room temperature C_V is close to $\frac{5}{2}R$. This discrepancy, one of the historically important failures of classical theory, is resolved by proper consideration of quantum effects as suggested by Einstein. The allowed energy levels of a harmonic oscillator are

not continuous as in classical mechanics, but are given by

$$\epsilon_n = nh\nu, \text{ with } n = 0, 1, 2, 3, \dots$$

where h is Planck's constant, ν is the natural frequency of the oscillator, and the energies ϵ_n are measured from the lowest level. For the average energy of a single oscillator, quantum statistical mechanics gives

$$\bar{\epsilon} = \frac{h\nu}{e^{h\nu/kT} - 1}$$

Instead of the classical value k , the contribution to C_V is

$$\frac{d\bar{\epsilon}}{dT} = k \frac{\left(\frac{h\nu}{kT}\right)^2 e^{h\nu/kT}}{(e^{h\nu/kT} - 1)^2} \quad (10)$$

This contribution to C_V is negligible for $T \ll 0.1 h\nu/k$ because most of the oscillators remain in the $n = 0$ level, but it increases rapidly near $T \approx h\nu/k$ and approaches the classical value for $T \gg h\nu/k$. For many diatomic molecules, $h\nu/k$ is a few thousand degrees Kelvin or more, and a room-temperature C_V of $\frac{5}{2}R$ can be understood as the sum of the translational and rotational contributions. The details of the temperature dependence predicted by Eq. (10), including the approach of C_V to $\frac{7}{2}R$ at sufficiently high temperatures have been verified experimentally for a number of gases.

In a few cases (H_2 , HD and D_2), diatomic gases exist at temperatures for which kT is comparable to the spacing of the rotational energy levels, and quantum effects can be observed in the rotational contribution to C_V . For H_2 , C_V drops below $\frac{5}{2}R$ as the temperature is reduced below about 300 K, and below 50 K it becomes equal to the $\frac{3}{2}R$ that is characteristic of translation. The translational energy levels of ordinary gases are so closely spaced that quantum effects are never important in the translational heat capacity.

The heat capacity of polyatomic gases can be treated by a straightforward generalization of the foregoing discussion. A molecule with n atoms has three translational degrees of freedom and, if it is nonlinear, three rotational and $3n - 6$ vibrational degrees of freedom. If it is linear, it has two rotational and $3n - 5$ vibrational degrees of freedom.

Solids. For every solid there is a lattice heat capacity associated with vibrations of the atoms. If the interatomic forces are harmonic, the N atoms in one mole of a monatomic solid have $3N$ independent vibrational modes, and the lattice heat capacity is the sum of $3N$ terms given by Eq. (10). Since the spectrum of the $3N$ frequencies of a real solid is complicated and difficult to calculate, an approximation introduced by Debye is widely used. In the Debye model, the vibrational modes are sound waves in an elastic continuum; the boundaries of the solid determine the allowed wavelengths and the sound velocities then determine the frequencies. To limit the number of frequencies to $3N$, the spectrum is cut off at a

maximum frequency $k\Theta_D/h$ where Θ_D , the Debye temperature, is determined by the sound velocities and is typically a few hundred degrees Kelvin. The cutoff corresponds to the fact that in a real crystal, the vibrations have a minimum wavelength comparable to the interatomic distance. The heat capacity is

$$C_V = \frac{12}{5} \pi^4 R \left(\frac{T}{\Theta_D} \right)^3 \quad (11)$$

for $T < \Theta_D/20$, but it increases less rapidly at higher temperatures and approaches the classical limit $3R$ for $T \rightarrow \Theta_D$. The predicted high-temperature limit is in agreement with the empirical rule of Dulong and Petit: for most monatomic solids C_V is approximately $3R$ at room temperature. However, this value is often exceeded at very high temperatures, partly as a consequence of anharmonicity in the interatomic forces. At low temperatures, Eq. (11) is in good agreement with experiment; C_V is found to be proportional to T^3 , although often only at temperatures below $\Theta_D/100$, and Θ_D is given accurately by the sound velocities. This agreement is to be expected because at low temperatures only low-frequency vibrations contribute and these *are* sound waves and therefore treated accurately in the Debye model. At intermediate temperatures, the agreement with experiment is only approximate.

Occasionally the lattice heat capacity of a molecular crystal is represented by the sum of a Debye heat capacity for the vibrations of the N molecules as units and the appropriate number of terms given by Eq. (10) for the intermolecular vibrations.

In metals, there is an electronic heat capacity related to the translational motion of the conduction electrons. The small mass and high density of the electrons make quantum effects important, and Sommerfeld showed that their heat capacity should be proportional to temperature for temperatures below about 10^4 K. This contribution is usually significant below a few degrees Kelvin, where the lattice heat capacity is small, and also at high temperatures, where it contributes to deviations from the rule of Dulong and Petit.

The heat capacities of a number of solids have bumps or peaks superimposed on the smoothly varying lattice and electronic heat capacities. These are usually called anomalies and two distinct types can be recognized. A Schottky anomaly is a smooth bump that arises from a set of energy levels for a single particle. The splitting of the rotational states of magnetic ions by electric fields, and of nuclei by electric or magnetic fields, are examples for which the associated anomalies occur in the ranges 10^{-1} to 10^3 K and 10^{-3} to 10^{-1} K, respectively. Lambda anomalies are sharp peaks produced by cooperative processes involving many particles in a transition from a low-temperature ordered state to a high-temperature disordered state. Examples are: the momentum ordering of ^4He atoms in liquid ^4He at 2.18 K and of electrons in superconductors at temperatures ranging from 10^{-1} to 10^4 K; the

magnetic ordering of ferromagnets and antiferromagnets at temperatures from 10^{-3} to 10^3 K; the spatial ordering of different atoms of an alloy on a superlattice, e.g., β -brass at 750°K; and the rotational disorder in certain molecular crystals e.g., H_2 at 1.5°K.

NORMAN E. PHILLIPS

Cross-references: CALORIMETRY, GAS LAWS, HEAT, TEMPERATURE AND THERMOMETRY, THERMODYNAMICS.

HEAT TRANSFER

Establishment of thermodynamic equilibrium for a system consisting of a number of media requires that the temperature be locally uniform and time-wise constant. A departure from this condition causes a transfer of energy in the form of heat from locations with high temperature to locations with low temperature. Such an energy transfer occurs very frequently and is encountered in our everyday life as well as in many engineering applications or in scientific experiments. It has, therefore, been known for a long time. The fact that quantitative predictions have become possible in the recent past only is due to the situation that several mechanisms are usually involved and interrelated in such an energy transfer process. They are generally classified as conduction, convection and radiation.

Heat transfer by conduction is that process which transports heat in a medium from one location to another without involvement of any visible movement. It is generally the only or the dominating mode of heat transfer in a solid medium; however, it occurs also in liquids and gases. In such fluids, this energy transport is often augmented when parts of the fluid are in movement and carry energy along. This mechanism of heat transfer is classified as transfer by convection. All media can also release energy in the form of photons (electromagnetic waves). This energy travels in space essentially with the velocity of light until the photons are recaptured by some other atoms, causing in this way heat transfer by radiation. An example of this energy transport is the transfer of heat from the sun to the earth. The three modes of energy transfer mentioned above will be discussed consecutively in the following sections. However, it must be realized that they often occur simultaneously, so that in some cases the total energy transport will be the sum of the contributions of the individual mechanisms. In other cases, such a summation will not lead to the correct result when the individual transport mechanisms mutually interfere.

Conduction. From a microscopic standpoint, thermal conduction refers to energy being handed down from one atom or molecule to the next one. In a liquid or gas, these particles change their position continuously even without visible movement and they transport energy also in this way. From a macroscopic or continuum viewpoint, thermal conduction is quantitatively described by Fourier's equation which states that the heat flux

q per unit time and unit area through an area element arbitrarily located in the medium is proportional to the drop in temperature, $-\text{grad } T$, per unit length in the direction normal to the surface and to a transport property k characteristic of the medium and called thermal conductivity.

$$q = -k \text{ grad } T \quad (1)$$

Predictions for the value of the thermal conductivity k can be made from considerations of the atomic structure. Accurate values, however, require experimentation in which the heat flux q and the temperature gradient $\text{grad } T$ are measured and these values are inserted into Fourier's equation. Figure 1 presents thermal conductivity values for a number of media in a large temperature range. It can be recognized that metals have the largest conductivities and, among those, pure metals have larger values than alloys. Gases, on the other hand, have very low heat conductivity values. Electrically nonconducting solids and liquids are arranged in between. The low thermal conductivity of air is utilized in the development of thermally insulating materials. Such materials, like cork or glass fiber, consist of a solid substance with a very large number of small spaces filled by air. The thermal transport occurs then essentially through the air spaces, and the solid structure only supplies the framework which prevents convective currents. The range of thermal conductivities in Fig. 1 at ambient temperature

extends through 5 powers of 10. This range is still small compared with the range for the electric conductivity of various substances where electric conductors have values which are larger by 25 powers of 10 than electric insulators. As a consequence of this fact, it is much easier to channel electricity along a desired path than to do so with heat, a fact which accounts for the difficulty in accurate experimentation in the field of heat transfer.

Fourier's equation can be used together with a statement on energy conservation to derive a differential equation describing the temperature field in a medium. Fourier was the first one to develop this equation and to devise means for its solution. In vector notation, this equation is

$$\rho c \frac{\partial T}{\partial t} = \nabla(k \nabla T) \quad (2)$$

where ρ is the density, c is the specific heat, t is time, ∇ is the Nabla operator. The temperature field in a substance can either change in time (unsteady state) or it can be independent of time (steady state, $\partial T / \partial t = 0$). For a steady-state situation, the temperature field depends primarily on the geometry of the body involved and on the boundary conditions. The simplest case of a steady state temperature field is the one in a plane wall with temperatures which are uniform on each surface, however different at the two surfaces. The temperature in the wall then changes linearly in the direction of the surface normal as long as the variation of the thermal conductivity in the temperature range involved can be neglected. For an unsteady process, the capacity of the medium to store energy enters the energy conservation equation; correspondingly, the specific heat of the material and its density become factors for the conduction process, as well as the thermal conductivity. A combination of these properties, defined as the ratio of the thermal conductivity to the product of specific heat and density, called

thermal diffusivity $\left(\frac{k}{\rho c}\right)$, then determines how fast existing temperature differences in a medium equalize in time. It is found that metals and gases have thermal diffusivity values which are approximately equal in magnitude and considerably higher than thermal diffusivities of liquid and solid nonconductors. This means that temperature differences equalize much faster in metals and gases than in other substances.

Various other physical processes lead in their mathematical description to equations of the same form as Eq. (2), especially in its steady-state form. Such processes are, for instance, the conduction of electricity in a conductor or the shape of a thin membrane stretched over a curved boundary. This situation has led to the development of analogies (electric analogy, soap film analogy) to heat conduction processes which are useful because they often offer the advantage of simpler experimentation.

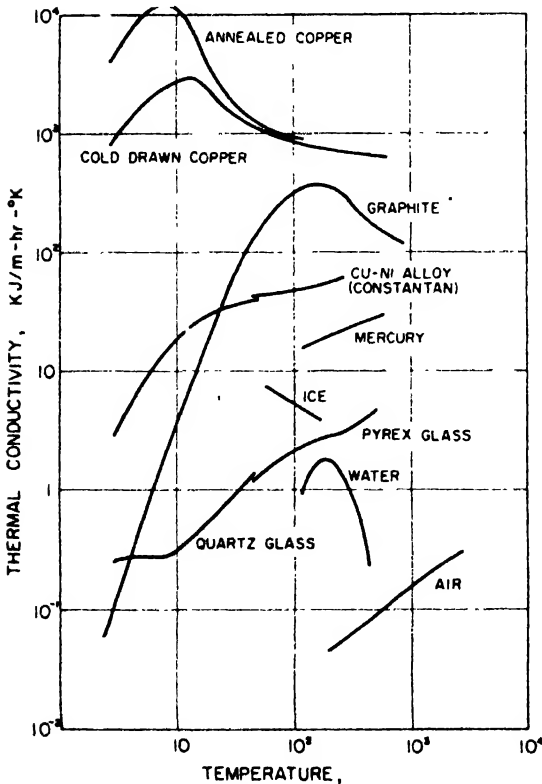


FIG. 1.

Convection. It has been mentioned before that in fluids, energy is often transported by convection. In such a situation, conduction takes care of the transport of heat from one stream tube to another and is the dominating mode of transfer near solid walls. Convection transports heat along the stream lines and is dominating in the main body of the fluid where the velocities are large. In many situations, the flow is turbulent; this means that unsteady mixing motions are superimposed on the mean flow. These mixing motions contribute also to a transport of heat between stream tubes, a process which can be thought of as being described by an "effective" conductivity which often has values by several powers of ten larger than the actual conductivity of the fluid.

The movement of the fluid may be generated by means external to the heat transfer process as by fans, blowers, or pumps. It may also be created by density differences connected with the heat transfer process itself. The first mode is called *forced convection*, the second one *natural or free convection*. Convective heat transfer may also be classified as heat transfer in duct flow or in external flow (over cylinders, spheres, air foils or similar subjects). In the second case, the heat transfer process is essentially concentrated in a thin fluid layer surrounding the object (boundary layer).

Of special interest in such heat transfer processes is the knowledge of the heat flux from the surface of a solid object exposed to the flow. This heat flux q_w per unit area and time is conventionally described by Newton's equation

$$q_w = h(T_w - T_f) \quad (3)$$

where T_w is the surface temperature and T_f is a characteristic temperature in the fluid. This equation defining the heat transfer coefficient h is convenient because in many situations the heat flux is at least approximately proportional to the temperature difference $T_w - T_f$. Information on the heat transfer coefficients can be obtained by a solution of the Navier-Stokes equations describing the flow of a viscous fluid and the related energy equation, or they are found by experimentation. The availability of electronic computers has tremendously increased our ability to study heat transfer analytically at least for laminar flow, whereas in turbulent flow the bulk of our information is based on experiments. Such experimentation is made difficult by the large number of parameters involved. Dimensional analysis has therefore been applied to reduce the number of influencing parameters, and relations for convective heat transfer are correspondingly presented in modern handbooks as relations between dimensionless parameters. Such an analysis demonstrates for instance, that heat transfer in forced flow can be described by a relation of the form

$$Nu = f(Re, Pr) \quad (4)$$

in which the Nusselt number Nu is a dimensionless parameter $\frac{hL}{k}$, containing the heat transfer

coefficient h ; the Reynolds number $Re = \frac{\rho V L}{\mu}$ describes essentially the nature of the flow; and the Prandtl number $Pr = \frac{c_p \mu}{k}$ can be considered as a dimensionless transport property characterizing the fluid involved. L and V are an arbitrarily selected characteristic length and velocity, respectively, ρ denotes the density, μ the viscosity, and c_p the specific heat of the fluid at constant pressure. Occasionally the Stanton number, $St = \frac{Nu}{Re Pr}$, is used instead of the Nusselt number

as a dimensionless expression of the heat transfer coefficient. Equation (4) is based on the assumption that the thermodynamic and transport properties involved in the heat transfer process can be considered as constant. Larger variations of such properties are usually accounted for by additional terms in Eq. (4) expressing the ratio of the varying transport properties or of parameters (temperature, pressure) on which they depend. Fluids occurring in nature cover a very large range of Prandtl numbers. Liquid metals, for instance, have values of order 0.001 to 0.01. Gases have values between 0.6 and 1, and oils have values up to 10 000 and more. Some heat transfer relations for forced convection are presented in Table 1. Relations for other situations can be found in the various texts mentioned at the end of this section or in corresponding handbooks. With regard to the relations in Table 1 and other similar equations, it has to be kept in mind that they are valid over a restricted range of the independent parameters only.

Heat transfer by free convection is described by relations of the form

$$Nu = f(Gr, Pr) \quad (5)$$

In this equation, the Grashof number

$$Gr = \frac{\rho^2 g \beta (T_w - T_f) L^3}{\mu^2}$$

(where g is the gravitational constant, β is the thermal expansion coefficient, T_f is the fluid temperature at a distance where it is not influenced by the heated or cooled object with surface temperature T_w) replaces the Reynolds number. Sometimes a dimensionless parameter called Rayleigh number ($Ra = Gr Pr$) is used instead of the Grashof number. Equation (5) assumes again that the fluid properties involved are nearly constant. Examples for such relations are also contained in Table 1. Examples for free convection situations are the heat transfer from a heating register in the room. Free convection is also an important factor in the establishment of the temperature in the atmosphere. In the free convection relations of Table 1 and in Eq. (5), it has been assumed that the convection flows are generated by the gravitational field. Natural convection can also be created by other body forces, like centrifugal and Coriolis forces or electromagnetic forces.

TABLE 1. RELATIONS FOR CONVECTIVE HEAT TRANSFER

*Forced Convection**Channel Flow*

Flow through a tube:

laminar ($Re < 3000$)

$$Nu = 3.65$$

turbulent ($Re > 3000, Pr > 0.6$)

$$Nu = 0.116 (Re^{2/3} - 125) Pr^{1/3}$$

$$\left(Nu = \frac{hD}{k}, Re = \frac{\rho D V}{\mu} \right)$$

 D - diameter, V - mean velocity
External Flow

Flat plate parallel to flow:

laminar ($1000 < Re < 500000, Pr > 0.6$)

$$Nu = 0.332 \sqrt{Re} \sqrt{Pr}$$

turbulent ($Re > 500000, Pr > 0.6$)

$$Nu = \frac{0.0297 Re^{4/5} Pr}{1 + 1.3 Re^{-1/4} Pr^{-1/4} (Pr - 1)}$$

$$\left(Nu = \frac{hx}{k}, Re = \frac{\rho V_\infty x}{\mu}, x = \text{distance from leading} \right)$$

 $\text{edge, } V_\infty = \text{velocity outside boundary layer}$
Cylinder normal to flow: $1 < Re < 4000$

$$Nu = 0.43 + 0.48 \sqrt{Re}$$

$$\left(Nu = \frac{hD}{k}, Re = \frac{\rho V_0 D}{\mu_1}, D = \text{diameter, } V_0 \right)$$

 upstream velocity
Natural Convection

Vertical flat plate:

laminar ($10^4 < Gr < 10^9$)

$$Nu = 0.508 Gr^{1/4} Pr^{1/2} (0.952 + Pr)^{-1/4}$$

turbulent ($Gr > 10^9$)

$$Nu = 0.0295 Gr^{2/5} Pr^{1/3} (1 + 0.494 Pr^{2/3})^{-1/5}$$

$$\left(Nu = \frac{hx}{k}, Gr = \frac{\rho^2 g \beta (T_w - T_f) x^3}{\mu^2}, x = \text{distance} \right)$$

 $\text{from leading edge, } T_w = \text{wall temperature, } T_f$
 $\text{fluid temperature at some distance from plate}$

(All surfaces are assumed to have uniform temperature)

Space does not permit the discussion of other heat transfer processes, although such processes have found increasing attention in recent years. Especially large heat transfer coefficients are created by a boiling or condensation process. Boiling heat transfer is therefore used in applications which have to deal with very large heat fluxes like chambers and nozzles of rockets or the anodes in electric arc devices. Heat transfer is also often combined with mass transfer processes. This is, for instance, the case in evaporation devices. Heat transfer may also occur combined with chemical reactions as in processes involving gases at very

high temperature where combustion, dissociation, or ionization occur.

Radiation. Energy can also be transferred from one location to another within a medium or from one medium to another in the form of photons (electromagnetic waves). Usually a multiplicity of wavelengths λ is involved in such energy transfer. In vacuum, all waves regardless of their wavelength move with the same speed $2.9977 \times 10^8 \text{ m/sec}$. In various substances, the wave velocity c changes somewhat with wavelength, and the ratio of the wave velocity in vacuum to the velocity in a substance is equal to the optical refraction index. Air and generally all gases have refraction indices which differ from one only in the fourth decimal. Their wave velocity is, therefore, practically equal to that in vacuum.

Prevost's principle states that the amount of energy emitted by a volume element within a radiating substance is completely independent of its surroundings. Whether the volume element increases or decreases its temperature by the process of radiation depends on whether it absorbs more foreign radiation than it emits or vice versa. One talks about thermal radiation when the emission of photons is thermally excited, i.e., when the substance within the volume element is nearly in thermodynamic equilibrium. The discussion in this section will be restricted to thermal radiation. For such radiation, Kirchhoff was able to derive a number of relations by consideration of a system of media in thermodynamic equilibrium. If j_ν indicates the co-efficient of emission, i.e., the radiative flux at the frequency ν emitted per unit volume into a unit solid angle, and α_ν is the co-efficient of absorption at the same frequency, that is, the fraction of the intensity of a radiant beam which is absorbed per unit path length, then one of these relations states

$$c^2 \frac{j_\nu}{\alpha_\nu} = f(T, \nu) \quad (6)$$

with c denoting the wave velocity. According to this relation, the combination of parameters on the left-hand side of the equation is a function of temperature T and frequency ν of the radiation only, but does not depend on the substance under consideration. Kirchhoff's law can also be expressed in parameters which refer to the interface of two media 1 and 2. It then takes the form

$$c^2 \frac{i_\nu}{\alpha_\nu} = f(T, \nu) \quad (7)$$

in which i_ν is the monochromatic intensity of the radiative flux at frequency ν originating in medium 2 and traveling through the interface into medium 1 per unit solid angle and area normal to the direction of the radiant beam. α_ν is the monochromatic absorptance or absorptivity, i.e. that fraction of a radiant beam approaching the interface in the medium 1 in the opposite direction that

* Frequency and wavelength are used interchangeably. They are connected by the relation $\lambda \nu = c$.

is absorbed in medium 2. c is the wave velocity in medium 1. Kirchhoff's law states that the combination of the parameters on the left-hand side of Eq. (7) is again a function of temperature and frequency only, but does not depend on the nature of the medium. A medium which absorbs all the radiation traveling into it through an interface ($\alpha_r = 1$) is called a blackbody. The intensity of radiation emitted by an arbitrary medium is, according to Eq. (7), in the following way related to the intensity of radiation i_{bb} emitted by a blackbody at the same temperature and frequency

$$\frac{i_\nu}{\alpha_r} = i_{bb} \quad (8)$$

From the consideration of a system in thermodynamic equilibrium, it is also easily shown that i_{bb} is independent of direction and that the total monochromatic radiant flux emitted by a blackbody per unit interface area and unit time is equal πi_{bb} .

The law describing the monochromatic intensity of radiation of a blackbody is given by Planck's equation

$$i = \frac{2h\nu^3}{c^2(e^{h\nu/kT} - 1)} \quad (9)$$

(where h is Planck's quantum constant and k is Boltzmann's constant). Figure 2 presents the wavelength dependence of the intensity of blackbody monochromatic radiation for a number of temperatures. It may be observed that for each temperature, the intensity has a maximum at a certain wavelength and that this maximum shifts toward short wavelengths with increasing temperature (Wien's law). The wavelength λ is plotted on the abscissa in microns μ ($1000\mu = 1$ mm). Our eye is sensitive to radiation in the range 0.4 to 0.7μ (the dashed range). It may be observed that for temperatures with which we

have largely to deal, the bulk of blackbody radiation is contained in the range of wavelengths larger than the visible ones (infrared range). This statement also holds for other media because Eq. (8) shows that no medium can have a monochromatic intensity which is higher than that of a blackbody. Only radiation coming from the sun has a major portion of the energy in the visible wavelength range (corresponding to a temperature of 6500°K).

The total energy flux q emitted per unit area and time from a blackbody can be obtained by integration of Eq. (9) over all frequencies and by multiplication of the result by π . For blackbody radiation into a vacuum (or with good approximation into a gas), the result is

$$q = \sigma T^4 \quad (10)$$

The Stefan-Boltzmann constant σ has the value $5.67 \times 10^{-8} \text{ W/m}^2 \text{ K}^{-4}$.

The following additional relation exists at an interface between two substances.

$$\rho_r + \alpha_r + \tau_r = 1 \quad (11)$$

This equation describes that the radiant energy in a beam approaching in a medium 1 the interface to a medium 2 is found again either as radiation reflected back into the medium 1 or absorbed in the medium 2 or transmitted through the medium 2 into other media or back into medium 1. The monochromatic reflectance or reflectivity ρ_r is the ratio of reflected to incident radiant energy, α_r is the corresponding ratio for the absorbed and τ_r for the transmitted energy. The vast majority of solids and liquids absorb radiant energy over most wavelengths in the infrared range within a very thin layer adjacent to the interface (of order 1μ to 1 mm). In heat transfer calculations, it can therefore usually be assumed that the transmissivity of such substances is equal to zero and that reflectivity and absorptivity are connected to the temperature of the interface. One talks then often in a simplified manner about radiative interchange between surfaces. Kirchhoff's law, Eq. (8), additionally connects the intensity of a beam emitted through the interface with the absorptivity and the intensity of a beam leaving a blackbody at the same temperature. Electromagnetic theory shows that electric nonconducting materials have generally high values of the absorptivity and correspondingly low values of the reflectivity. Metals (electric conductors), on the other hand, behave in the opposite way, having low absorptivity and high reflectivity values. This fact is utilized in aluminum-insulations and in vacuum thermos bottles. In the visible range, the appearance of surfaces of various materials to the eye already supplies information on approximate values of the reflectivity and absorptivity. A white surface, for instance, has a very low absorptivity. Gases behave differently with respect to radiation. They need fairly large layers in order to absorb the major part of incident radiation and radiate only in restricted wavelength ranges whereas solids and liquids have a more continuous spectrum.

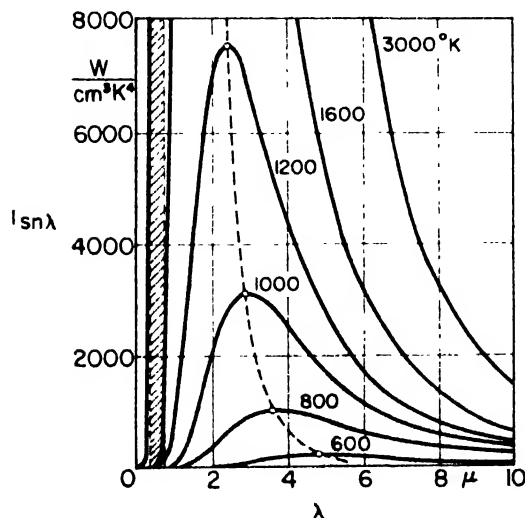


FIG. 2.

Values for the radiation properties (ρ_r , α_r , τ_r) together with the relations for blackbody radiation are the basis of calculations to determine heat exchange by radiation in a system with locally varying temperature. Calculations of such interchange are in general very involved, especially when substances with small absorption coefficients are involved. The formulation of such interchange leads to integral differential equations. The reader has in this connection to be referred to the books listed at the end of this section, and only a few relations for simple geometries will be presented here. Consider two area elements, dA_1 and dA_2 , belonging to two blackbodies with the temperatures T_1 and T_2 . The distance between the two area elements is s , and β_1 and β_2 are the angles between the two surface normals and the interconnecting line. The following equation then describes the net heat transfer dQ from area element dA_1 to area element dA_2 per unit time assuming that no radiation is absorbed or emitted in the space between the two surfaces

$$dQ = \frac{\cos \beta_1 \cos \beta_2}{\pi s^2} dA_1 dA_2 \sigma (T_1^4 - T_2^4) \quad (12)$$

If non-black surfaces are involved, then the process of radiant interchange is much more involved, since part of the incident radiation is now reflected from the surfaces and travels in this way back and forth until it is finally absorbed. Simple relations exist in this case for the radiative interchange between the surfaces of two concentric spheres or cylinders with areas A_1 and A_2 and temperatures T_1 and T_2 . It is further assumed that both surfaces are emitting and reflecting in a perfectly diffuse way and that they are separated by a medium which does neither emit nor absorb radiation. The net monochromatic interchange $d\Phi_r$ between the two surfaces is then described by the equation

$$d\Phi_r = \frac{\pi}{\frac{1}{\alpha_{r1}} + \frac{A_1}{A_2} \left(\frac{1}{\alpha_{r2}} - 1 \right)} dA_1 (i_{br1} - i_{br2}) \quad (13)$$

The monochromatic intensities i_{br1} and i_{br2} are calculated with Eq. (9). The relation changes when the outer cylinder reflects radiation mirror-like (specularly) to

$$d\Phi_r = \frac{\pi}{\frac{1}{\alpha_{r1}} + \left(\frac{1}{\alpha_{r2}} - 1 \right)} dA_1 (i_{br1} - i_{br2}) \quad (14)$$

Both equations merge asymptotically into the same relation when the differences between the two radii become small. The corresponding relation then also holds for two parallel infinite planes. Equations (13) and (14) have to be integrated over all frequencies to obtain the net heat transfer between the two surfaces. The result is simple when the absorptances α_{r1} and α_{r2} are independent of frequency (gray surfaces). Equations (13) and (14) describe then the net heat

transfer, when i_{br1} and i_{br2} are replaced by $\frac{\sigma}{\pi} T_1^4$ and $\frac{\sigma}{\pi} T_2^4$, respectively.

E. R. G. ECKERT

References

- Schneider, P., "Conduction Heat Transfer," Reading Mass. Addison-Wesley Publishing Co., 1955.
 McAdams, W. H., "Heat Transmission," Third edition, New York, McGraw-Hill Book Co., 1954.
 Jakob, M., "Heat Transfer," Vols. I and II, New York, John Wiley & Sons, 1949 and 1957.
 Eckert, E. R. G., and Drake, R. M., Jr., "Heat and Mass Transfer, Second edition, New York, McGraw-Hill Book Co., 1959.

Cross-references: HEAT, HEAT CAPACITY, INFRARED RADIATION, REFLECTION.

HEISENBERG UNCERTAINTY PRINCIPLE

Classical physics is based on two assumptions that have been found experimentally to be untenable. The first of these is the existence of signals that can travel with infinite speed; the second is that the magnitude of the interaction between two systems can be reduced to arbitrarily small values. The realization that the speed of propagation of signals has a finite upper limit led to the development of relativity theory. The recognition of the existence of a finite quantum of action has been incorporated in quantum (wave) mechanics.

Quantum mechanics assigns a physical reality only to those variables whose value can, in principle, be experimentally determined. About the existence of phenomena or systems that cannot be experimentally observed, quantum mechanics is noncommittal. Questions regarding an isolated system are meaningless in quantum mechanics, for any observation made on such a system necessarily disturbs its isolation by at least one quantum of action. Heisenberg⁵ observed that any measurement made on a system destroys some of the knowledge gained about that system through previous measurements. Any prediction about the future course of a system must be contingent on a knowledge about the measurements that will be made on that system, and is subject to uncertainties introduced by the measurements. Whereas one might speculate with some reliability about the future course of a system under the restriction that no more measurements will be performed on it, such speculations would be physically meaningless, as they could not be experimentally confirmed or denied.

In the broadest sense, then, the Heisenberg Uncertainty Principle states that the partitioning of the universe into observer (either a human observer, or a recording device such as a photographic plate) and observed is subject to a finite inaccuracy; one might say that the "knife" or "pencil" that makes the partition has a finite "thickness", h .

The concept of a monochromatic beam of radiation is not a difficult one to accept. Yet the experimental determination of the frequency of such a beam requires an infinite time interval; any finite portion chopped from the beam is shown by Fourier analysis to have a spectrum of finite width, hence not a single frequency at all. Quantum mechanics does not deny the existence of a monochromatic beam, but it does render the assignment of a definite frequency in a given time interval meaningless.

Fourier analysis shows that the specifications of the time interval, Δt , and of the spectral width, $\Delta\nu$, are reciprocally related:

$$\Delta\nu \Delta t \geq 1 \quad (1)$$

This equation represents the uncertainty relation for classical waves: any attempt to specify the frequency at an instant of time results in a broadening of the frequency spectrum. This uncertainty relation applies to any wave, whether electromagnetic, acoustic, or otherwise.

Interference patterns observed when electron beams are reflected from crystalline surfaces³ or transmitted through thin metallic films⁷ indicate that these beams possess some wave characteristics. DeBroglie², by independent theoretical considerations (see WAVE MECHANICS), postulated the following relations between the dynamic variables, energy (E) and linear momentum (p) of the beam, on the one hand, and the wave variables, frequency (ν) and wavelength (λ), on the other:

$$E = h\nu \quad (2)$$

$$p = h/\lambda \quad (3)$$

Substitution of Eq. (2) in Eq. (1) gives the uncertainty relation between energy and time:

$$\Delta E \Delta t \geq h \quad (4)$$

For a beam of free electrons of mass m , traveling in the x direction, $\Delta t = m \Delta x/p_x$, where p_x is the linear momentum and Δx is the distance traveled in the time interval Δt . Since $E = p_x^2/2m$, $\Delta E \cong p_x \Delta p_x/m$, so that $\Delta E \Delta t = \Delta p_x \Delta x$, and

$$\Delta p_x \Delta x \geq h \quad (5)$$

The pairs of variables (E, t) and (p_x, x) are called canonically conjugate pairs of variables. In quantum mechanics, the operators corresponding to canonically conjugate variables do not commute (see QUANTUM THEORY). Heisenberg originally stated his uncertainty principle in the following form: the values of canonically conjugate variables of a given system can only be determined with a finite lower limit of accuracy.

Among the many important experimental phenomena illustrating the uncertainty principle is the COMPTON EFFECT. Here, a photon is scattered by an electron; the momentum of the photon is rendered uncertain as a result of its scattering by the electron, and the electron is moved from its original position by the impact received from

the photon. If we consider the photon and the electron as separate systems, then their interaction (the collision) introduces uncertainties for each system, given by Equation (5).

A general uncertainty relation follows from the postulates of WAVE MECHANICS. Consider two variables, a and b , of a system, whose operators are \mathcal{A} and \mathcal{B} . The expectation values of a and b are called \bar{a} and \bar{b} respectively. The uncertainties in a and b can then be defined quantitatively as their respective rms deviations from their expectation values; it follows from the postulates that

$$(\Delta a)^2 \equiv \oint \Psi^* (\mathcal{A} - \bar{a})^2 \Psi dq$$

where Ψ is the normalized wave function of the system under observation, and $\oint \dots dq$ indicates integration over all values of all coordinates. Similarly,

$$(\Delta b)^2 \equiv \oint \Psi^* (\mathcal{B} - \bar{b})^2 \Psi dq$$

If \mathcal{A} and \mathcal{B} are Hermitian,[†] $(\mathcal{A} - \bar{a})^2$ and $(\mathcal{B} - \bar{b})^2$ are also Hermitian. Therefore:

$$\begin{aligned} (\Delta a)^2 \cdot (\Delta b)^2 &= \oint \Psi^* (\mathcal{A} - \bar{a})^2 \Psi dq \cdot \oint \Psi^* (\mathcal{B} - \bar{b})^2 \Psi dq \\ &= \oint (\mathcal{A} - \bar{a}) \Psi (\mathcal{A} - \bar{a})^* \Psi^* dq \\ &\quad \cdot \oint (\mathcal{B} - \bar{b}) \Psi (\mathcal{B} - \bar{b})^* \Psi^* dq \\ &= \oint |(\mathcal{A} - \bar{a}) \Psi|^2 dq \cdot \oint |(\mathcal{B} - \bar{b}) \Psi|^2 dq. \end{aligned}$$

To put this product of two definite integrals in a more useful form, consider the function

$$f(q) = \lambda(\mathcal{A} - \bar{a}) \Psi + i(\mathcal{B} - \bar{b}) \Psi^*$$

where λ is real, and independent of coordinates, and $i \equiv \sqrt{-1}$. The function $|f(q)|^2$ must be non-negative:

$$\begin{aligned} \lambda^2 [(\mathcal{A} - \bar{a}) \Psi]^2 &- i\lambda [(\mathcal{A}^* - \bar{a}) \Psi^* (\mathcal{B} - \bar{b}) \Psi + \\ &+ (\mathcal{A} - \bar{a}) \Psi (\mathcal{B}^* - \bar{b}) \Psi^*] + |(\mathcal{B} - \bar{b}) \Psi|^2 \geq 0 \end{aligned}$$

When the left-hand side of this inequality is integrated over all values of all coordinates:

$$\begin{aligned} \lambda^2 (\Delta a)^2 + (\Delta b)^2 &+ i\lambda [\oint (\mathcal{A}^* - \bar{a}) \Psi^* (\mathcal{B} - \bar{b}) \Psi dq + \\ &- \oint (\mathcal{A} - \bar{a}) \Psi (\mathcal{B}^* - \bar{b}) \Psi^* dq] \geq 0 \end{aligned}$$

Since λ is real, the left-hand side of this inequality becomes negative unless the discriminant becomes zero or negative:

$$\begin{aligned} 4(\Delta a)^2 (\Delta b)^2 &\geq -[\oint (\mathcal{A}^* - \bar{a}) \Psi^* (\mathcal{B} - \bar{b}) \Psi dq + \\ &- \oint (\mathcal{A} - \bar{a}) \Psi (\mathcal{B}^* - \bar{b}) \Psi^* dq]^2 \end{aligned}$$

[†] An operator \mathcal{F} is Hermitian, if for any properly behaved functions u and v : $\oint u \mathcal{F} v dq = \oint v \mathcal{F}^* u dq$, where $*$ indicates complex conjugation. Quantum mechanical operators usually are, or can be made to be, Hermitian.

The right-hand side of this inequality is reduced as follows (remember that Ψ is normalized):

$$\begin{aligned} & \oint (\mathcal{A}^* - \bar{a}) \Psi^* (\mathcal{B} - b) \Psi dq - \oint (\mathcal{A} \Psi)^* \mathcal{B} \Psi dq + \\ & \quad - \bar{a} \oint \Psi^* \mathcal{B} \Psi dq - b \oint \Psi^* (\mathcal{A} \Psi)^* dq + \bar{a} b \\ & \oint (\mathcal{A} - \bar{a}) \Psi (\mathcal{B}^* - \bar{b}) \Psi^* dq - \oint (\mathcal{A} \Psi) (\mathcal{B} \Psi)^* dq + \\ & \quad - \bar{a} \oint \Psi (\mathcal{B} \Psi)^* dq - b \oint \Psi^* (\mathcal{A} \Psi) dq + \bar{a} b \\ \text{Since } b &= \oint \Psi^* \mathcal{B} \Psi dq = \oint \Psi^* (\mathcal{B} \Psi)^* dq, \text{ and} \\ \bar{a} &= \oint \Psi^* \mathcal{A} \Psi dq = \oint \Psi^* (\mathcal{A} \Psi)^* dq, \\ 4(\Delta a)^2 (\Delta b)^2 &\geq - [\oint (\mathcal{A} \Psi)^* (\mathcal{B} \Psi) dq + \\ & \quad - \oint (\mathcal{A} \Psi) (\mathcal{B} \Psi)^* dq]^2 \end{aligned}$$

Since \mathcal{A} and \mathcal{B} are Hermitian

$$\begin{aligned} \oint (\mathcal{A} \Psi)^* (\mathcal{B} \Psi) dq &\equiv \oint (\mathcal{B} \Psi) (\mathcal{A} \Psi)^* dq = \\ &= \oint \Psi^* \mathcal{A} \mathcal{B} \Psi dq, \\ \oint (\mathcal{A} \Psi) (\mathcal{B} \Psi)^* dq &= \oint \Psi^* \mathcal{B} \mathcal{A} \Psi dq \quad (6) \\ \therefore 4(\Delta a)^2 (\Delta b)^2 &\geq - [\oint \Psi^* (\mathcal{A} \mathcal{B} - \mathcal{B} \mathcal{A}) \Psi dq]^2 \\ \text{and } |\Delta a| |\Delta b| &\geq \frac{1}{2} \oint \Psi^* (\mathcal{A} \mathcal{B} - \mathcal{B} \mathcal{A}) \Psi dq \end{aligned}$$

According to the postulates of wave mechanics, the operators for momentum and position are given respectively by: $\mathcal{P} = -i\hbar \nabla$, where $\hbar = h/2\pi$, and $\mathcal{Q} = q$ (multiplication by q).

For linear motion in the x direction, $\mathcal{P}_x = -i\hbar \partial/\partial x$, $\mathcal{Q} = x$. Hence $(\mathcal{P}\mathcal{Q} - \mathcal{Q}\mathcal{P})\Psi = -i\hbar (c/\partial x)(x\Psi) = i\hbar x(\partial\Psi/\partial x) = i\hbar\Psi$. When this expression is substituted into Eq. (6), it follows that:

$$|\Delta p_x| |\Delta x| \geq \hbar/4\pi$$

which is in agreement with Eq. (5). Equation (4) can be similarly derived from the postulates of wave mechanics by setting $\mathcal{E} = (i\hbar/2\pi)(\partial/\partial t)$, $\mathcal{T} = t$.

A. L. LOEB

References

- Bohm, David, "Quantum Theory," Englewood Cliffs, N.J. Prentice-Hall, 1951.
- De Broglie, L., *J. Phys. Ser. 6*, 7, 1 "Introduction to Wave Mechanics," London, Methuen, (1926).
- Davison, C. J., and Germer, L. H., *Phys. Rev.*, **30**, 705 (1927); *Proc. Natl. Acad. Sci. U.S.*, **14** 317 (1928).
- Harris, L., and Loeb, A. L., "Introduction to Wave Mechanics," New York, McGraw-Hill Book Co., 1963.
- Heisenberg, W., *Z. Physik*, **43**, 172; "The Physical Principles of the Quantum Theory," Chicago, University of Chicago Press, (1927).
- Margenau, H., and Murphy, G. M., "The Mathematics of Physics and Chemistry," Princeton, N.J., D. Van Nostrand, 1943.
- Thomson, G. P., *Proc. Roy. Soc. London Ser. A*, **117** 600 (1928); **119**, 651 (1928).

Cross-references: COMPTON EFFECT, MATRIX MECHANICS, QUANTUM THEORY, WAVE MECHANICS.

HIGH-VOLTAGE RESEARCH

High-voltage research deals with phenomena evoked by high voltages and intense electric fields, with the behavior of dielectrics and electrical components under such electrical stress, and with the utilization of electrostatic fields and forces for various purposes of science and industry. In the laboratory, this electrical stress is produced by the presence of electric charge on the opposing surfaces of two electrodes between which a voltage is applied. Pulsed, alternating and constant voltages ranging from a few kilovolts to 10 MV have been used in such studies. In nature, air currents and water precipitation cause electric charge to become separated between cloud and earth or between clouds; the stressed region may reach electrical pressure differences in excess of 1000 MV. Lightning, a rapid high-current discharge, completes the breakdown of the over-stressed air and dissipates the accumulated electrical energy. Many aspects of this natural high-voltage phenomena have been the subject of investigation because of their scientific interest and the danger of lightning to life and to susceptible structures such as electric power systems.

In the industrial high-voltage laboratory, the direct and induced effects of lightning discharges on electrical apparatus are often simulated by high-voltage impulse generators. These use the Marx method of first slowly charging a number of condensers in parallel and then suddenly connecting them in series by spark-gap switches which at the same instant impress the multiplied voltage upon the test circuit. A typical voltage wave produced by such impulse generators rises to its peak value of several million volts in $1 \mu\text{sec}$ and then diminishes exponentially reaching half-voltage in $10 \mu\text{sec}$.

Electric Field between High-voltage Electrodes. The region between and around electrodes which have been charged by the application of a voltage V between them is occupied by the electric field of that charge. In an isolated system, the positive electrode has a deficiency of electrons exactly equal to the excess electrons on the negative electrode. The amount of electric charge Q on either electrode surface at any instant is given by $Q = CV$ where Q is in coulombs and C is the capacitance of the electrode system in farads. Energy is stored in the electrically stressed space between the electrodes; the amount of this electrical energy W can be expressed in terms of applied voltage or separated charge by $W = CV^2/2 = Q^2/2C$ joules.

The electric field intensity at any point is defined by the magnitude and direction of the force which would be experienced by a unit positive charge placed in the field at that point. Following Faraday, if we define a line of electric flux as a line drawn so that its direction is everywhere the direction of the force on a positive particle, and require that one line must originate on each unit positive charge and terminate on each unit negative charge, then the lines of flux

will map out the electric field and the lines per unit area will be directly related to the electrical field intensity E . The electrostatic force acting on a particle with charge q placed in an electric field of intensity E volts per meter is given by $f = qE$ newtons. The electric field distribution depends upon the geometry of the electrodes but is affected by the presence of dielectric materials or of charged particles; a quantitative picture of the static or changing field picture is usually essential to high-voltage research.

Objectives of High-voltage Research. From antiquity to the present, the history of high-voltage research sparkles with many names well-known in electrical science—Thales of Miletus, von Guericke, Newton, Franklin whose kite experiment established the identity of natural lightning and electricity, Cavendish, Faraday, and Roentgen whose discovery of x-rays in 1895 marked the beginning of the atomic age. During the 60 years centered on the turn of the twentieth century, physicists studied the passage of electricity through gases at normal and reduced pressure, sought an understanding of long sparks and corona in atmospheric air, measured the conductivity and breakdown strength of liquid dielectrics, and examined the flashover of solid insulators in these media. More recent research has been concerned with the electrical strength and dielectric losses of many solid dielectrics from porcelain to hydrocarbon polymers, the increasing voltage-insulating ability of gases with pressures extending to scores of atmospheres, and the mechanisms which lead to electrical conduction and breakdown between electrodes immersed in high vacuum.

Research in the high-voltage field may also be directed at testing and increasing the insulation strength of power equipment subject to lightning or switching surges. The increasing trend toward higher voltages for the transmission of electric power over long distances has directed research toward reduction of radio interference and power loss by corona and surface leakage from high-voltage transmission line conductors and their suspension insulators. The need to bring such power into urban areas has produced the requirement for ac power cables in which the center conductor is reliably insulated for hundreds of kilovolts above earth. The inherent efficiency and stability of dc power transmission have led to the development of more adequate high-voltage rectification and conversion apparatus. For these purposes, the low-pressure metallic vapor tube has reached the highest power levels though solid state devices offer much promise for high-voltage, high-power switching.

In the field of science and medicine, high-voltage research seeks improved methods of producing high constant voltages, of measuring and stabilizing such voltages, and of applying them to the acceleration of atomic ions and electrons to high energies. Such particle accelerators are needed for nuclear structure research; for the study of the properties of energetic atoms, electrons, x-rays, and neutrons; for the treatment

of deep-seated tumors with energetic particles and radiation, and for the radiation processing of materials.

Industrial objectives of high-voltage research include the development of high-power, high-frequency tubes and their power supplies for radar and long-range radio communication, the ionization of particulate matter by corona and its electrostatic collection as in smoke and chemical precipitators, and the elimination of electrostatic hazards which arise in processes such as the transfer or mixing of volatile hydrocarbons, dusts, and explosive gases.

JOHN G. TRUMP

References

- Graggs, J. D., and Meek, J. M., "High Voltage Laboratory Technique," Butterworths Scientific Publications, London, 1954.
- Trump, J. G., "Electrostatic Sources of Electric Power," *Elec. Eng.*, **66**, No. 6, 525-534 (June 1947).
- Trump, J. G., and Van de Graaff, R. J., "The Insulation of High Voltages in Vacuum," *J. Appl. Phys.*, **18**, 327-332 (March 1947).

Cross-references: ELECTRICITY, POTENTIAL, ACCELERATOR, VAN DE GRAAFF.

HISTORY OF PHYSICS

Sir J. J. Thomson, discoverer of the electron, in his address on the nature of light, at Cambridge, England in June 1925, summarized very well the most important years in the growth and evolution of physical science. He stated that "thirty years ago I would have little to say about the nature of light whereas today it is the story of the entire physics." Thomson evidently referred to that dramatic period when the triumphant rise of nineteenth century physics reached its climax in the last decade of that century, with several revolutionary discoveries that symptomatized a new era. The nineteenth century represents such an overwhelming milestone in the unfolding process of man's knowledge of the physical world and its laws that in a brief account it is bound to dwarf all other periods into relatively fleeting time intervals. As a mighty mountain top, nineteenth century physics alone offers an eloquent panorama of the place and role of scientific discipline in the history of civilization.

Physics, as we know it today, is a relatively young science. As an independent field, it had its actual beginning with the age of Renaissance. Then distinctly associated with the pioneering work of Galileo, it was firmly established by Newton and his mathematization of the processes of physical ideas. Fragmentary rudiments of physics can, of course, be traced to the famous fountainheads of western civilization in the ancient river civilizations of Egypt and Mesopotamia and above all to their clearing house, ancient Greece. Whatever physics there was, it had its sustenance in a close association with astronomy, or astrology as it

was then known, man's truly oldest science. Its adepts at that time were philosophers rather than scientists. Having some portentous fear of and aversion toward experiment, they were timid observers of the majestic natural phenomena. Out of the ages of mythological mentality, Socratic ratiocination emerged in the ancient Greek miracle, and in the complex maze of natural phenomena tried to "save the appearances." This was achieved through philosophical speculation. Aristotle, a giant in pure philosophy, also wrote considerably on physics, yet his erroneous ideas on mechanics, due to his undisputed authority for many centuries, played a negative if not paralyzing role in the field of physical science. Some less-known thinkers such as Democritus remained in the background, although with their sound ideas on the nature of matter and space they could have had a strongly catalytic influence on the subsequent growth of physics.

The physical science of antiquity can be summed up in a most appropriate way by reference to the ancient Alexandrian school or library. This library, as a form of academy, symbolizes amazing accomplishments where almost all major fields of science were more or less apprehended in their rudiments or fundamentals. This center, shining like its lighthouse, Pharos—one of the seven wonders of the world—represents the integration of all knowledge on the nature of world "physis" as gathered since the famous Pythagorean school up to the climax attained in this Greek colonial city in Egypt. It includes a galaxy of names which left permanent traces in the formative state of physical science. Euclides' *Stoicheia* and Ptolemy's *Almagest* are probably its most precious heritage.

The age of antiquity where the appearance of natural science is associated with the first rise of rational thought in the world of mythology came to an end with the decline of the Alexandrian school. This occurred within two centuries, following the destruction of the library first by the Christians in 390 A.D. and in 642 A.D. by the Mohammedans. A millenium of the medieval era followed, described usually as the "dark ages." Physics, which demands objectivity, had little chance in face of all the pervading and forceful sacerdotal propaganda. Nevertheless, the historian's truth even there was evident: When it grows dark enough, stars come out. After its initial iconoclastic zeal, Islam reversed its attitude toward ancient learning and some of its scholars carried very high the torch of impartial studies while the Christian western world's concern was the salvation of the soul. Thus, the works of the ancient Greek philosophers, that encouraged interest in physical science, first reached Europe in translation from Arabic texts. Two church fathers, however, prove that all was not darkness in this age. St. Augustine, Bishop of Hippo, and St. Thomas Aquinas with their penetrating ideas stood at the fringes of such physics that responded to the dream of medieval man.

The all-powerful ecclesiastic rulers could not escape their decadence, and a change of tide was

inevitable. Whether it was the ecumenical meeting in Florence, in 1438, with Byzantine and Roman delegates gathered from both factions of the divided Christian world, or whether it was the life and work of that most unusual genius, Leonardo da Vinci, 1452-1519, the new stream of Renaissance relentlessly heralded an unprecedented era. Above all, it was the science of physics which, with the pioneers of astronomy, was to start its expansion destined to change the face of the world. Leonardo da Vinci in Italy and Nicholas Copernicus in Polish Pomerania (1473-1543) are probably most responsible for the wave of the future in which physics obtained an impulse as never before in history. As contemporaries, both were born still deep within the medieval society in widely different parts of Europe, yet their work corresponded and their accomplishments are firmly established in the foundations of classical physics.

The actual beginning of physics, however, is undoubtedly associated with Galileo Galilei (1564-1642). Standing at the cradle of physics, Galileo confronted a herculean task befitting his versatile genius. His mind displayed an inexhaustible industry and imagination when he took up the defense of the Copernican heliocentric system into the realm of the physical laboratory. Very few of his opponents were aware that through his ingenious demonstrations, such as rolling balls along an inclined plane, he would build lasting foundations of modern mechanics. It is rather curious that a man, most popular for his telescope which he did not invent, has gained undying glory for the introduction of modern scientific method, based upon experimentation. Galilean science is synonymous with empiricism and aposterioristic truth, i.e., that which is acquired by experience. Thus, Galileo as a pioneer of a new science experienced the full impact of opposition by his contemporaries, the ruling Scholastic doctrinaires, who advocated the dogma of aprioristic truth represented by Aristotelian speculative philosophy. It was at this stage that Aristotle's exponents misused his philosophy to the detriment of the healthy advancement of physics. The planetary currents heralding the new age, however, were invincible.

While Galileo died under house arrest and in disgrace, and that enthusiastic defender of Copernicanism, J. Kepler, died in misery, all glory and recognition were heaped upon Isaac Newton during his lifetime. It is significant that he was born the year Galileo passed away, 1642. The notable inscription on Newton's tombstone in Westminster Abbey characterizes the gratitude of his nation "Mortals should congratulate themselves that such a spiritual knight was born amongst them." Yet Newton was modest enough to give credit to the giants on "whose shoulders he was sitting," as he once described his own work. *Principia*, his opus major, stands out among the few greatest products of the human intellect. We use the term Galilean science, but classical physics is referred to as Newtonian physics, associated with the world view of the Newtonian

universe. This was considered the final state of perfection of physics whose advancement, as Albert Michelson commented, would consist only in the position of the decimal point.

Newtonian physics, formulated in three laws of motion and climaxed by the universal law of gravitation, continued in its progressive refinement in order to reach its triumphant culmination in the nineteenth century. Its primary hallmark is the mathematization of space. Mathematical physics was described by Imanuel Kant as the greatest achievement of the human intellect. Indeed, it had a most fascinating confirmation in the mathematical discovery of the planet Neptune by J. C. Adams and Leverrier, in 1845. Laplace produced an overwhelming impression on the entire century when, in 1798, he used the rigor of mathematics to formulate his genial hypothesis on the origin of the solar system. The traditional heritage of the association of astronomy with physics is alive with Galileo and Newton, for both were astronomers and physicists. From that time on, the gap widened until now we have physicists who are almost complete strangers to astronomy.

With the culminating period of fruition, Newtonian physics is signalized by legions of great names as builders of our view of the physical world. It is possible here to note only a few principal milestones, each a giant of his own in a panoramic view of the glorious century. From Laplace at the beginning of the nineteenth century, the epic unfolds from Avogadro to Faraday, from Carnot to Helmholtz and Kelvin, attaining its peak in Maxwellian equations formulating the electromagnetic theory of light. These equations were impressively described by the physicist, Boltzmann, when he quoted Goethe's Faust: "Who was the god who wrote these lines?" Yet, these mathematical equations that so well described the natural processes, pointed inevitably to a deterministic universe. It was also a mechanistic universe. Thus, the nineteenth century physicist sought to create a model of such a universe. The corpuscular-kinetic nature of the universe, as seen by the physicist of the Victorian age, appeared only as a huge machine. It was a universe that offered no mysteries that could not be resolved in due course of time. Ultimately the universe was knowable and predictable. This *weltanschauung* (world view) of triumphant physics encouraged the rise of materialistic philosophy that was actually shaped by Marx and Engels into dialectical materialism, destined to become the official doctrine of the ruling communist state of the twentieth century.

Newtonian physics was not destined to remain the last form of knowledge. The turn of the twentieth century once again witnessed another tidal wave that changed the course of the evolution of physics. Whereas society at present lives by the impact of physics produced by its rise in the previous century, the revolutionary discoveries

that occurred in physics around 1895-1905 contain unfathomable, portentous consequences for the future. Becquerel's radioactivity, Roentgen's x-rays, Thomson's electron, Planck's quantum, Einstein's relativity, represent a revolution that will carry man incomparably farther into the mysteries of the universe than the Copernican revolution ever did. Heisenberg's principle of indeterminacy in the realm of microphysics not only discards the once cherished Laplacean determinism but points to the existence of a universe incomparably more complex than indicated by Newtonian mechanism. "It is not the discovery of an outlying island," stated Sir J. J. Thomson, then President of the Royal Society at the announcement of the first empirical confirmation of Einsteinian relativity in 1919, "it is a whole continent of new scientific ideas. It is the greatest discovery in connection with gravitation since Newton enunciated his principle."

Yet, it was still a quarter of a century before an even greater dramatic confirmation of the Einsteinian formula of the equality of energy with the product of mass and the square of the velocity of light, which marked, with a great human tragedy, the beginning of the age of nuclear energy. It is this aspect of modern, nuclear physics which through the atomic microcosmos reunites the physicist with the astrophysicist. The latter, in turn, investigates stellar laboratories in the macrocosmos where matter is subjected to conditions inaccessible to any terrestrial laboratories. This teamwork, in addition to the supreme geometrization of space-time of relativistic physics is bound to unveil new wonders of the heretofore unimaginable Einsteinian universe. Yet Einstein once humbly retorted to a eulogy on his great discoveries: "All I did was to study the history of physics."

KAREL HUJER

References

- Sarton, George, "A History of Science," Cambridge, Mass., Harvard University Press, 1952, 2 vols.
- Shapley, H., Wright, H., and Rapport, S., "Readings in Physical Science," New York, Appleton, Century, Crofts, 1948.
- Magie, W. F., "A Source Book in Physics," 1st. Ed., McGraw-Hill Book Co., New York, 1935.
- Jeans, James, "Growth of Physical Science," London, Oxford University Press, 1948.
- Omer, Knowless, Mundy, and Yoho, "Physical Science, Men and Concepts," Boston, 1962.
- Taylor, F. Sherwood, "A Short History of Science and Scientific Thought," New York, 1949.
- Price, Derek J. de S., "Science Since Babylon," London, 1962.

HYDRODYNAMICS, See FLUID DYNAMICS.

I

IMPULSE AND MOMENTUM

The concept of impulse and momentum derives directly from Newton's law of motion.

Consider first the case of *linear* impulse and *linear* momentum. Newton's law states that in the proper frame of reference, force is equal to the (time) rate of change of momentum, where momentum is defined as the product of mass and velocity. Consider a force F acting for a time Δt on a particle of mass m , thereby changing its velocity from v_1 to v_2 . The rate of change of momentum in this case is the change of momentum divided by the time interval during which the change occurs. Thus, Newton's law states

$$F = \frac{(mv)_2 - (mv)_1}{\Delta t}$$

Now if we multiply each side of this equation by Δt , we obtain

$$F \Delta t = (mv)_2 - (mv)_1$$

The left-hand side of this equation, $F \Delta t$, representing the product of a force and the time interval through which the force acts, is called "impulse." Thus, this equation states what is often known as the *Law of Impulse and Momentum*: Impulse is equal to the change of momentum.

It becomes apparent that a body will experience the same change of momentum irrespective of the separate values of F and Δt as long as their product $F \Delta t$ is the same. Thus, a large force acting briefly may have the same net effect as a smaller force acting longer, if the two impulses are the same. Going to the limit, we may consider an infinite force acting for an infinitesimal time such that their product remains a finite quantity. Under the action of such an impulse, a body will experience an instantaneous change of velocity. A common example is the change of velocity of a baseball as it is hit by a bat. For all practical purposes the change occurs instantaneously.

As a consequence of the Law of Impulse and Momentum described above, we find that if the total force is zero, so is the total impulse, and the momentum will remain unchanged. This is known as the *Conservation Law of Momentum*. It applies either to one particle or to a system of particles. Consider, for instance, the collision of two billiard balls, A and B. Taking each ball separately, the impulse on A by B is equal in magnitude but opposite in direction to the impulse on B by A. Thus, upon collision, the change of momentum of

A is also equal and opposite to that of B. On the other hand, if we take *both* A and B as our system, then the two impulses are acting on the same system. Since the two are equal and opposite, they cancel out and the total impulse on this system is zero. The Conservation Law of Momentum then predicts that the total momentum of the system (which is the sum of momenta of all particles in the system) remains a constant no matter what goes on in the system.

The Conservation Law of Momentum forms one of the basic cornerstones in physics and engineering. Its application is all-pervading, from the motion of stars to the encounter and scattering of molecules, atoms and electrons.

We will now continue our discussion at a more precise level. First of all, in calculating the momentum mv , the mass m should be the relativistic mass defined as

$$m = \frac{m_0}{\sqrt{1 - (v/c)^2}}$$

where m_0 is the rest mass, i.e., mass at zero velocity, and c is the velocity of light. It is seen that at velocities much less than the velocity of light, the relativistic mass and the rest mass are indistinguishable. Secondly, both force and velocity are vector quantities, i.e., they have a magnitude and a direction and obey the parallelogram law of addition, and we shall use letters \mathbf{F} and \mathbf{v} to represent them. Finally, as the time interval Δt approaches zero as a limit, the rate of change of momentum during Δt becomes a derivative and Newton's law becomes:

$$\mathbf{F} = \frac{d}{dt}(m\mathbf{v})$$

After integrating each side of this equation with respect to time t and taking the integration limits from $t = t_1$ to $t = t_2$, the result is:

$$\int_{t_1}^{t_2} \mathbf{F} dt = (m\mathbf{v})_2 - (m\mathbf{v})_1$$

The integral on the left-hand side is called "impulse," and we again reach the Law of Impulse and Momentum given previously. In evaluating this integral, we must know the variation \mathbf{F} as a function of time, i.e., $\mathbf{F} = \mathbf{F}(t)$. Furthermore, in applying this law to a system of particles, we need only to count those impulses that are caused by *external* forces acting on the

particles due to sources outside the system. As has been illustrated by the previous example on two billiard balls, the *internal* forces that any two particles exert on each other will generally cancel out if both particles are included in the system.

Up to this point we have discussed the laws of linear impulse and linear momentum. Entirely similar laws hold for *angular* impulse and *angular* momentum. The angular momentum of a particle about a point O is defined as

$$\mathbf{L} = \mathbf{r} \times m\mathbf{v}$$

where \mathbf{r} is the radius vector from O to the particle. The moment of force or *torque* about O is defined as

$$\mathbf{T} = \mathbf{r} \times \mathbf{F}$$

If we take the cross product of \mathbf{r} with each side of the expression for Newton's law as given before, we obtain

$$\mathbf{r} \times \mathbf{F} = \mathbf{T} = \mathbf{r} \times \frac{d}{dt}(m\mathbf{v})$$

The right-hand side can be identified to be just $d\mathbf{L}/dt$ on account of the vector identity:

$$\frac{d\mathbf{L}}{dt} = \frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) = \mathbf{v} \times m\mathbf{v} + \mathbf{r} \times \frac{d}{dt}(m\mathbf{v})$$

where the first term on the right-hand side vanishes. Thus,

$$\mathbf{T} = \frac{d\mathbf{L}}{dt}$$

We now integrate this equation with respect to t and take the integration limits from $t = t_1$ to $t = t_2$. The result is

$$\int_{t_1}^{t_2} \mathbf{T} dt = \mathbf{L}_2 - \mathbf{L}_1$$

Analogous to the linear impulse, we may call the integral on the left-hand side of the above equation the angular impulse. Thus this equation states: Angular impulse is equal to the change of angular momentum. In particular, if the total torque is zero, so is the total angular impulse, and the angular momentum will remain unchanged. This is known as the *Conservation Law of Angular Momentum*.

HSUAN YFH

References

- Yeh, Hsuan, and Abrams, Joel I., "Principles of Mechanics of Solids and Fluids," Vol. I, New York, McGraw-Hill Book Co., 1960.
 Synge, John L., and Griffith, Byron A., "Principles of Mechanics," New York, McGraw-Hill Book Co., 1949.
 Goldstein, Herbert, "Classical Mechanics," Reading, Mass., Addison-Wesley, 1950.
Cross-references: CONSERVATION LAWS AND SYMMETRY, DYNAMICS, MECHANICS, ROTATION-- CIRCULAR MOTION, STATICS, VECTOR PHYSICS.

INDUCED ELECTROMOTIVE FORCE

Electromotive force and voltage drop are usually regarded as synonymous. When an electromotive force, or simply emf, is impressed on a closed metallic circuit, current results. The emf along a specified path C in space is defined as the work per unit charge W/q done by the electromagnetic fields on a small test charge moved along C . Since work is the line integral of force \mathbf{F} , the work per unit charge is the line integral of the force per unit charge. Letting \mathbf{F}/q denote the vector electromagnetic force per unit charge in newtons per coulomb, we have

$$\text{emf} = \int_C \frac{\mathbf{F}}{q} \cdot d\mathbf{l} \text{ volts} \quad (1)$$

The scalar product $(\mathbf{F}/q) \cdot d\mathbf{l}$ is the product $(F/q) \cos \theta dl$, with θ denoting the angle between the vectors \mathbf{F}/q and $d\mathbf{l}$.

The electric force per unit charge is the electric field intensity \mathbf{E} (volts per meter) and the magnetic force per unit charge is $\mathbf{v} \times \mathbf{B}$, with \mathbf{v} denoting the velocity of the test charge in meters per second and \mathbf{B} denoting the magnetic flux density in webers per square meter. In terms of the smaller angle θ between \mathbf{v} and \mathbf{B} , the cross product $\mathbf{v} \times \mathbf{B}$ is a vector having magnitude $vB \sin \theta$; the direction of the vector $\mathbf{v} \times \mathbf{B}$ is normal to the plane of the vectors \mathbf{v} and \mathbf{B} , with the sense of that of the extended thumb of the right hand oriented so that its fingers curl through the angle θ from \mathbf{v} toward \mathbf{B} . As the total force per unit charge is $\mathbf{E} + \mathbf{v} \times \mathbf{B}$, the emf in terms of the fields is

$$\text{emf} = \int_C (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l} \quad (2)$$

It might appear from Eq. (2) that the emf depends on the forward velocity with which the test charge is moved along the path C . However, this is not the case. If \mathbf{v} and $d\mathbf{l}$ in Eq. (2) have the same direction, then the vectors $(\mathbf{v} \times \mathbf{B})$ and $d\mathbf{l}$ are normal, and their scalar product is zero. Consequently, only the component of \mathbf{v} normal to $d\mathbf{l}$ can contribute to the emf. This component has value only if the differential path length $d\mathbf{l}$ has sideways motion. Thus \mathbf{v} in Eq. (2) represents the *sideways motion, if any, of $d\mathbf{l}$* . The fields \mathbf{E} and \mathbf{B} of Eq. (2) may be functions of time as well as functions of the space coordinates. In addition, the velocity \mathbf{v} of each differential path length $d\mathbf{l}$ may vary with time. However, Eq. (2) correctly expresses the emf, or voltage drop, along the path C as a function of time. That component of the emf consisting of the line integral of $\mathbf{v} \times \mathbf{B}$ is known as the *motional emf*, because it has value only when the path C is moving through a magnetic field, cutting lines of magnetic flux. For stationary paths there is no motional emf, and the voltage drop is simply the line integral of the electric field \mathbf{E} .

For an emf to exist along a stationary path, it is necessary to have an electric field present. As electric charges are surrounded by electric fields emfs are generated by devices that separate

charge. A familiar example is the battery, which utilizes chemical forces to separate charge. Some other methods of separating charge are the heating of a thermocouple, the exposure of a photocell to incident light, and the rubbing together of different materials. Electric fields are also produced by time-changing magnetic fields, and this principle is extensively exploited, as is motional emf, to generate electric power. The remainder of this article is devoted to electromotive force induced by magnetic means.

A fundamental law of electromagnetism, often called the *Maxwell-Faraday law*, or the *first law of electromagnetic induction*, states that the line integral of the electric field intensity \mathbf{E} around any closed path C equals $-\partial\phi/\partial t$, with ϕ representing the magnetic flux over any surface S having the closed path C as its contour. The positive side of the surface S and the direction of the line integral around the contour C are related by the right-hand rule; by this rule, the curled fingers are oriented so as to point around the loop in the direction of the integration and the extended thumb points out of the positive side of the surface S . The magnetic flux ϕ is the surface integral of the magnetic flux density \mathbf{B} ; that is,

$$\phi = \iint_S \mathbf{B} \cdot d\mathbf{S} \text{ webers} \quad (3)$$

In Eq. (3) the vector differential surface $d\mathbf{S}$ has area dS and is directed normal to the plane of $d\mathbf{S}$ out of the positive side. The partial time derivative of ϕ is defined as

$$\frac{\partial\phi}{\partial t} = \iint_S \frac{\partial\mathbf{B}}{\partial t} \cdot d\mathbf{S} \text{ volts} \quad (4)$$

and this is often referred to as the *magnetic current* through the surface S . For a moving surface S the limits of the surface integral of Eq. (4) are functions of time, but Eq. (4) still applies. It is important to understand that in evaluating $\partial\phi/\partial t$ over a surface that is moving in a region containing a magnetic field, we treat the surface at the instant under consideration as though it is stationary. The partial time derivative of ϕ is the time rate of increase of the flux over the surface S due only to the changing magnetic field \mathbf{B} ; any increase in ϕ due to the motion of the surface in the \mathbf{B} -field is *not* included. The Maxwell-Faraday law is

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\frac{\partial\phi}{\partial t} \quad (5)$$

with ϕ being the magnetic flux in webers out of the positive side of any surface having the path C as its contour. The small circle on the integral sign indicates a closed path. We note from Eq. (5) that an electric field must be present in any region containing a time-changing magnetic field.

The application of Eq. (2) to a closed path C gives

$$\text{emf} = \oint_C \mathbf{E} \cdot d\mathbf{l} + \oint_C (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l} \quad (6)$$

Utilizing Eq. (5) enables us to write Eq. (6) in the form

$$\text{emf} = -\frac{\partial\phi}{\partial t} + \oint_C (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l} \quad (7)$$

Thus the emf around a closed path consists, in general, of two components. The component $-\partial\phi/\partial t$ is often referred to as the *variational emf* (or *transformer emf*), and the second component is, of course, the motional emf.

In Eq. (7) the relation $(\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l}$ can, by means of a common vector identity, be replaced with $\mathbf{B} \cdot (\mathbf{v} \times d\mathbf{l})$. As \mathbf{v} is the sideways velocity of $d\mathbf{l}$, the vector $\mathbf{v} \times d\mathbf{l}$ has magnitude $v dl$ and direction normal to the differential surface dS swept out by the moving length $d\mathbf{l}$ in the time dt . Letting B_n denote the component of \mathbf{B} normal to this area, we note that $\mathbf{B} \cdot (\mathbf{v} \times d\mathbf{l})$ becomes $B_n v dl$, and Eq. (7) can be written

$$\text{emf} = -\left[\frac{\partial\phi}{\partial t} + \oint_C B_n v dl \right] \quad (8)$$

Clearly the integral of $B_n v$ around the closed contour C , with t denoting the magnitude of the sideways velocity of each $d\mathbf{l}$, is simply the time rate of increase of the magnetic flux over the surface bounded by C due to the path C cutting lines of magnetic flux. Hence, the complete expression in brackets is the time rate of increase of the magnetic flux ϕ , over any surface S bounded by the closed path C , due to the changing magnetic field and also due to the moving path cutting through the magnetic field. Equation (8) is often written

$$\text{emf} = -\frac{d\phi}{dt} \quad (9)$$

It is important to note carefully the distinction between Eqs. (5) and (9). Equation (5) is only the variational emf, and Eq. (9) is the sum of the variational and motional emfs. In Eq. (5), the partial time derivative of the magnetic flux ϕ is the rate of change of the flux due only to the time-changing magnetic field; in Eq. (9), the total time derivative is the rate of change of the flux due to the time-changing \mathbf{B} -field and also to the path cutting through the magnetic field. Of course, if the closed path is not cutting lines of magnetic flux, then Eqs. (5) and (9) are equivalent. It is also important to note that $d\phi/dt$ in Eq. (9) does not necessarily mean the total time rate of change of the flux ϕ over the surface S . For example, the flux over a surface S bounded by the closed contour C of the left-hand electric circuit of Fig. 1 is changing when the coil is being unwound by the rotation of the cylinder. However, as \mathbf{B} is static there is no variational emf, and since the conductors are not cutting flux lines, there is no motional emf. Consequently, $d\phi/dt$ as used in Eq. (9) is zero even though the flux is changing with time. Note that $d\phi/dt$ in Eq. (9) was defined as representing the bracketed expression of Eq. (8), and $d\phi/dt$ must not be more broadly interpreted.

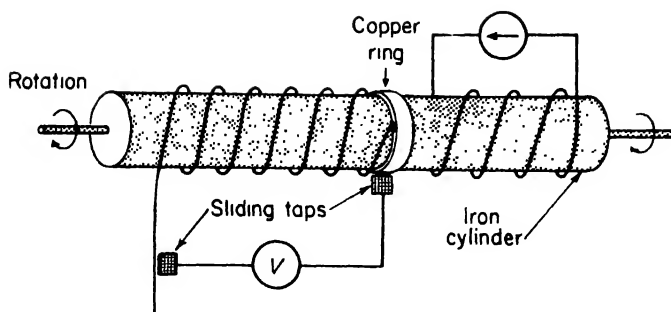


FIG. 1. The current generator produces a steady magnetic flux in the iron cylinder, which rotates as the wire is pulled at A .

In the applications of the equations which have been presented, we must refer all flux densities and movements to a single specified coordinate system. In particular, the velocities are with respect to this system and are not relative velocities between conductors and moving lines of flux. Of course the coordinate system is arbitrarily selected, and *the relative magnitudes of the variational and motional emfs depend upon the selection*. Let us consider two examples.

Example 1. An electric generator is shown in Fig. 2. The parallel stationary conductors separated a distance l have a stationary voltmeter connected between them. The electric circuit is completed through a moving conductor that is connected electrically by means of sliding taps. This conductor is at $y = 0$ at time $t = 0$ and moves to the right with constant velocity $v = v\mathbf{a}_y$. The applied flux density \mathbf{B} , represented in Fig. 2 by dots, is $B_0 \cos \beta y \cos \omega t \mathbf{a}_z$. Unit vectors in the directions of the respective coordinate axes are \mathbf{a}_x , \mathbf{a}_y , and \mathbf{a}_z . Find the instantaneous voltage across the voltmeter.

Solution. Let S denote the plane rectangular surface bounded by the closed electric circuit, with the positive side selected as the side facing the reader. The counterclockwise emf around the electric circuit is $-\frac{d\phi}{dt}$, with ϕ signifying the magnetic flux out of the positive side of S . As $dS = l dy \mathbf{a}_x$, the scalar

product $\mathbf{B} \cdot d\mathbf{S}$ is $B_0 l \cos \beta y \cos \omega t dy$; integrating from $y = 0$ to $y = y_1$ gives

$$\phi = (B_0 l / \beta) \sin \beta y_1 \cos \omega t \quad (10)$$

with y_1 denoting the instantaneous y -position of the moving wire. The counterclockwise emf is found by replacing y_1 with vt and evaluating $-\frac{d\phi}{dt}$. The result is

$$\text{emf} = \omega B_0 l / \beta \sin \beta vt \sin \omega t = B_0 l v \cos \beta vt \cos \omega t \quad (11)$$

The variational (transformer) component is $-\frac{d\phi}{dt}$, which is determined with the aid of Eq. (10) to be $\omega B_0 l / \beta \sin \beta y_1 \sin \omega t$, with $y_1 = vt$. This is the first component on the right side of Eq. (11). Note that y_1 was treated as constant when evaluating the partial time derivative of ϕ . The motional emf is the line integral of $\mathbf{v} \times \mathbf{B}$ along the path of the moving conductor. As $\mathbf{v} \times \mathbf{B}$ is $-B_0 v \cos \beta y_1 \cos \omega t \mathbf{a}_z$ and as $d\mathbf{l}$ is $dz \mathbf{a}_z$, we evaluate the integral of $-B_0 v \cos \beta y_1 \cos \omega t dz$ from $z = 0$ to $z = l$, obtaining a motional emf of $-B_0 l v \cos \beta y_1 \cos \omega t$. This component results from the cutting of lines of magnetic flux by the moving conductor.

If the voltmeter draws no current, there can be no electromagnetic force on the free electrons of the wires. Therefore, *the emf along the path of the metal conductors, including the moving conductor, is zero*. The total voltage of Eq. (11) appears across the voltmeter.

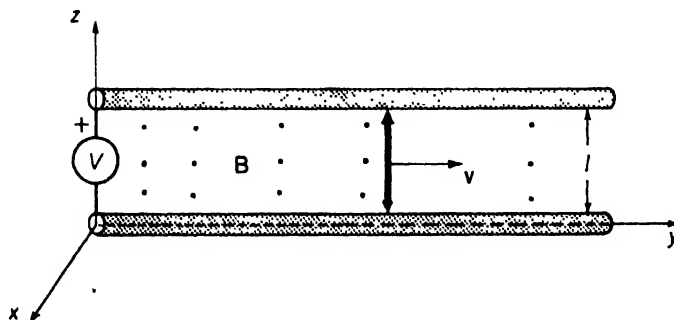


FIG. 2. Elementary electric generator.

Example 2. Suppose the conductor with sliding taps in Fig. 2 is stationary ($v = 0$) and located at $y = y_1$. Also suppose that the magnetic field \mathbf{B} is produced by a system of steady currents in conductors (not shown in Fig. 2) that are moving with constant velocity $\mathbf{v} = v\mathbf{a}_y$. At time $t = 0$ the magnetic field \mathbf{B} is $B_0 \sin \beta y \mathbf{a}_x$. Determine the voltage across the voltmeter.

Solution. There is no motional emf because the conductors of Fig. 2 are stationary with respect to our selected coordinate system. However, the magnetic field at points fixed with respect to the coordinate system is changing with time, and hence there is a variational emf.

As the \mathbf{B} -field at $t = 0$ is $B_0 \sin \beta y \mathbf{a}_x$ and moving with velocity $v\mathbf{a}_y$, the \mathbf{B} -field as a function of time is $B_0 \sin [\beta(y - vt)] \mathbf{a}_x$. This is verified by noting that an observer at y_0 at $t = 0$ moving in the y -direction with the velocity v of the moving current-carrying conductors, would have a y -coordinate of $y_0 + vt$; hence according to the expression for \mathbf{B} , he would observe a constant field. The magnetic current density is

$$\partial \mathbf{B} / \partial t = -\beta v B_0 \cos \beta(y - vt) \mathbf{a}_x$$

The negative of the integral of this over the rectangular surface bounded by the electric circuit, with the positive side selected as the side facing the reader and with y limits of zero and y_1 , gives the counter-clockwise emf. The result is

$$\text{emf} = B_0 v [\sin \beta(y_1 - vt) + \sin \beta vt]$$

This is the voltage across the meter.

CHARLES A. HOLT

References

- Bewley, L. V., "Flux Linkages and Electromagnetic Induction," New York, Dover Publications, 1964.
 Fano, R. M., Chu, L. J., and Adler, R. B., "Electromagnetic Fields, Energy, and Forces," New York, John Wiley and Sons, 1960.
 Holt, C. A., "Introduction to Electromagnetic Fields and Waves, New York, John Wiley & Sons, 1963.
 Moon, P., and Spencer, D. E., "Foundations of Electrodynamics," Princeton, N. J., D. Van Nostrand Co., 1960.

Cross-references: ALTERNATING CURRENTS, CIRCUITRY, INDUCTANCE, POTENTIAL.

INDUCTANCE

Inductance is a ratio of a magnetic flux Φ to an electric current i . The unit is the henry; 1 henry \equiv 1 weber/ampere. The *mutual inductance* M of two circuits is defined as the ratio of the magnetic flux, linking with one circuit, to the current in the other

$$M_{12} = \frac{\Phi_2}{i_1}, M_{21} = \frac{\Phi_1}{i_2}$$

The *self-inductance* L of a single circuit is defined

as the ratio of the flux linking the circuit to the current flowing in the circuit.

$$L = \frac{\Phi}{i}$$

In the absence of any magnetic material, M and L depend only on the geometry of the circuits concerned. The mutual inductance of two circuits 1 and 2 can be calculated from the Neumann equation*

$$M_{12} = M_{21} = \frac{\mu_0}{4\pi} \oint_1 \oint_2 \frac{\mathbf{dl}_1 \cdot \mathbf{dl}_2}{r}$$

where \mathbf{dl}_1 and \mathbf{dl}_2 are vector elements of length in circuits 1 and 2, respectively, and r is the distance between these two elements. Note that this expression is completely symmetrical with respect to the two circuits. The self-inductance is calculated from the same equation, where the elements \mathbf{dl}_1 and \mathbf{dl}_2 are now situated on the same circuit. The double integral is evaluated by first keeping the position of \mathbf{dl}_2 fixed and integrating \mathbf{dl}_1/r around the circuit. The process is then repeated for all other elements such as \mathbf{dl}_2 , and the results are summed.

The *external inductance* of a circuit is the part of its self-inductance which is due to flux lying outside the surface of the conductor while the *internal inductance* is the contribution from the magnetic flux within the conductor itself.

If the current, and therefore the magnetic flux associated with it, varies with time, an electromotive force (emf) will be induced in any circuit linked by the flux. This provides an alternative method of defining inductance in terms of the emf induced by a given rate of change of current.

$$\mathcal{E}_2 = -M \frac{di_1}{dt}$$

$$\mathcal{E}_1 = -M \frac{di_2}{dt}$$

$$\mathcal{E} = -L \frac{di}{dt}$$

The negative signs are used to imply that the direction of the emf is always such as to tend to oppose the change of current (Lenz's law). Note, however, that although the self-inductance must always be positive, the mutual inductance of two circuits may be either positive or negative. An inductance of 1 henry corresponds to an induced emf of 1 volt for a rate of change of current of 1 ampere/sec.

In order to maintain a current i , the induced emf $-L di/dt$ must be balanced by an equal and opposite applied voltage so that the total applied voltage is

$$v = Ri + L \frac{di}{dt}$$

* μ_0 is the permeability of free space; $\mu_0 = 4\pi \times 10^{-7}$ henrys/m.

and the power supplied to the circuit is

$$vi = Ri^2 + \frac{d}{dt} \left(\frac{1}{2} Li^2 \right)$$

where the resistance R is a measure of the power dissipated. This equation shows another way of interpreting inductance, namely as a measure of the amount of energy stored in the magnetic field when a given current flows. The stored energy is $\frac{1}{2} Li^2$ joules when L is measured in henrys and i in amperes.

Electric circuit theory is based on the use of sinusoidally varying currents and voltages. When the current varies sinusoidally with time, the rate of change of current has the same time-waveform except for a phase shift of $\pi/2$ radians. This leads to the idea of a complex impedance, of which the real part is associated with the power lost from the circuit and the imaginary part is a reactance given by the ratio of the magnitude of the induced emf to that of the current. Since the phase of the applied voltage leads that of the current, the magnetically induced reactance is taken as being positive. We thus have

$$i = I \sin \omega t$$

$$v = Ri + L \frac{di}{dt}$$

The circuit impedance is Z where

$$Z = \frac{v}{i} = R + j\omega L$$

The self-inductance of a circuit can be regarded as a parameter which determines the inductive reactance presented to a sinusoidally varying current of given amplitude and frequency. In the same way, the mutual inductance determines the mutual reactance between two circuits.

Owing to magnetic hysteresis, the voltage and current-time waveforms cannot both be sinusoidal if any magnetic material is present. Under these conditions, the reactance is defined as the ratio of the fundamental-frequency components of the voltage and current waveforms.

The sign of the mutual inductance can be specified in the following way. Suppose that the two circuits are connected in series. Then the total induced emf will be

$$-(L_1 + L_2 + 2M) \frac{di}{dt}$$

and the total reactance will therefore be

$$j\omega(L_1 + L_2 + 2M)$$

and will be greater than the sum of the reactances of the separate circuits when M is positive. A combination of two coils is said to be series aiding or series opposing, depending on whether they are connected so that M is positive or negative, respectively.

The Q -factor is used in describing the properties of an inductor. It is defined as

$$\frac{\omega L}{R} = \tan \phi$$

where ϕ is the phase angle of the complex impedance Z . An alternative term is the *power-factor* $\cos \phi$, particularly when the emphasis is on power dissipated rather than on the damping of tuned circuits.

An impedance Z can be represented by an equivalent circuit consisting of a resistance R in series with an ideal *series inductance* L . An alternative is to start with the admittance $Y = 1/Z$ and to represent this by the parallel combination of a resistance R' and a *parallel inductance* L' .

$$R = \frac{R'}{1 + Q^2}$$

$$L = L' \left(\frac{Q^2}{1 + Q^2} \right)$$

Note that, when Q is sufficiently large,

$$L' \simeq L$$

$$R' \simeq Q^2 R$$

Inductance Coil (Inductor). This is a device which is specially designed to possess inductance. The winding may have a ferromagnetic core composed of a dust-core material, metallic laminations or ferrite. An *air-cored* coil is one with no magnetic core.

Although it is the inductance which is of interest the influence of the electric field often cannot be neglected and causes the effect known as self-capacitance. At relatively low frequencies, the inductor will behave as if an equivalent lumped capacitance were shunted across its terminals. This simple equivalent circuit fails as the frequency is approached at which the apparent lumped capacitance would resonate with the inductance of the coil. The problem then becomes one of electromagnetic wave propagation, and the concept of inductance is no longer relevant. The effect of a fixed parallel capacitance C is to reduce the Q -factor to Q_c , where

$$Q_c = Q \left| 1 - \omega^2 LC \left(1 + \frac{1}{Q^2} \right) \right|$$

When Q is large, this becomes

$$Q_c \simeq Q(1 - \omega^2 LC)$$

The self-resonance effect is sometimes an advantage, e.g., in choke coils, where the object is to obtain a high impedance, irrespective of its phase angle. Generally however self-capacitance is an undesirable property, particularly in electrical networks in which the inductor forms part of a series resonant circuit.

The self-capacitance of a coil can be kept to a minimum by spacing the turns of the winding well

apart and also, in the case of a multi-layer winding, by ensuring that wires which lie physically close together always belong to adjacent parts of the winding so that the potential difference between them is relatively small.

The inductance of some types of coil can be calculated directly, without recourse to the Neumann equation. An example is the *toroidal coil*, i.e., one consisting of a uniform winding around a ring-shaped former. In this case the magnetic flux exists in a well-defined magnetic circuit and the inductance is given, very nearly, by

$$L = \frac{\mu_0 N^2 A}{l}$$

where N is the number of turns in the winding and A and l are, respectively, the cross-sectional area and the length of the magnetic circuit. The same formula also holds for a long, thin *solenoidal coil*. Corrections must be applied unless both the diameter of each turn and the radial depth of the winding are small compared with the length of the magnetic circuit for the toroid or the length of the coil itself for the solenoid.

If the toroidal former is replaced by one made of a magnetic material having a relative permeability μ , the inductance will be increased by a factor μ . For coils of other geometrical shapes, where some of the flux linking the winding may lie partly or wholly outside the magnetic core, the presence of the latter will increase the inductance by a factor called the *effective permeability* μ_e . This can never exceed μ and may be considerably smaller than μ .

The relative permeability of a ferromagnetic material is not a constant but is a function of the instantaneous flux density. This increases power losses and causes distortion of the current and voltage time-waveforms.

Power dissipation in a metallic core can be represented approximately by an equivalent series resistance R , where

$$\frac{R}{\mu_e f L} = c + h B_{\max} + ef$$

and B_{\max} is the peak flux density. The parameters c and h depend on the hysteresis properties of the core material. The parameter e is a measure of eddy-current losses in the core.

Ferrite materials are practically insulators; therefore eddy-current losses are negligible. For a ferrite, however, the relative permeability must be regarded as a complex quantity

$$\mu = \mu'(1 - j \tan \delta)$$

Both μ' and the dissipation coefficients $\tan \delta$ are functions of frequency.

The useful frequency range of coils with laminated cores is often limited by magnetic skin effect. This causes both the inductance and the Q -factor to start to fall as the depth of penetration of the electromagnetic field becomes comparable with the thickness of the laminations.

The depth of penetration d is given by

$$\frac{2}{d^2} = \omega \mu_0 \mu \sigma$$

where σ is the conductivity of the core material.

The inductance of a coil with a magnetic core is reduced if there is a superimposed unidirectional polarizing flux, caused for example by a dc current in the winding. This may be an unwanted effect or it may be used as a means of controlling the inductance.

Transverse air gaps are often introduced into the magnetic circuit. There are several reasons for this. An air gap reduces the flux density for a given magnetizing force, thus reducing the inductance of a given coil. The power losses are, however, reduced at a greater rate than the inductance so that the power factor is improved. The performance of the inductor is made less dependent on the magnetic parameters of the core material. The waveform distortion due to hysteresis is also reduced. An air gap may be used to prevent saturation of the core by the polarizing flux when the coil is required to carry a dc current.

If the inductance of a coil, with a closed magnetic circuit of length l , is L , an air gap of length g will reduce the inductance to L_g where

$$L_g = L \cdot \left[\frac{\mu}{1 + \mu(g/l)} \right]$$

As the gap ratio g/l is increased, the inductance tends to become independent of μ , particularly when μ is large. Note however that the formula assumes that the presence of the air gap does not change the geometry of the flux distribution; an assumption which is generally only justified for relatively small gap ratios.

Dust-cored or ferrite-cored coils, for which the core eddy-current losses are small, are often wound with stranded wire. This is done to keep the eddy-current losses in the winding low by ensuring that the individual conductor strands have diameters which are small compared with the depth of penetration of the electromagnetic field into the material of which they are made. At very high frequencies, where this condition cannot be met, solid wire is often used and the diameter of the wire is increased to compensate for the fact that only part of its cross-sectional area is effective as a conductor.

V. G. WELSBY

References

- Scott, W. T., "The Physics of Electricity and Magnetism," New York, John Wiley & Sons, 1959.
- Plonsey, R., and Collins, R. E., "Principles and Applications of Electromagnetic Fields," New York, McGraw-Hill Book Co., 1961.
- Welsby, V. G., "The Theory and Design of Inductance Coils," London, Macdonald & Co., 2nd Edition, 1960.

Cross-references: INDUCED ELECTROMOTIVE FORCE, TRANSFORMER.

INERTIAL GUIDANCE*

Many modern spacecraft, aircraft, ships, and submarines are being designed to navigate by inertial guidance systems which use only sensed acceleration and vehicle turning rates for input information. Early inertial guidance systems, developed by the Germans during World War II, simply gyro-stabilized the airframe to the desired flight attitude, and used a single accelerometer to measure acceleration along the longitudinal (thrust) axis. When the integrated acceleration reached the desired injection velocity, the engines were cut off. Many unsophisticated systems still use this method.

Principles of Operation. Conventional inertial systems today are designed to provide a gyro-stabilized platform which is gimbal-mounted to permit unlimited vehicle motion without disturbing the stable element. On the stable element are mounted linear accelerometers to measure the two or three components of the vehicle's acceleration vector.

These components of acceleration are inputs to the computer (Fig. 1*), which solves the navigation equations - adding computed gravitation, integrating to find velocity, and integrating again to determine position:

$$\mathbf{R} = \int_0^t \int_0^t (\mathbf{A} + \mathbf{G}) dt^2 + \mathbf{V}_0 t + \mathbf{R}_0$$

where

\mathbf{R} - position vector

\mathbf{A} - non-gravitational acceleration vector (sensed acceleration)

* Illustrations from Parvin's "Inertial Navigation" (Principles of Guided Missile Design Series), Princeton, N.J., D. Van Nostrand Company, Inc., 1962.

\mathbf{G} - gravitational vector (calculated)

\mathbf{V}_0 = initial velocity vector (inserted)

t = time

\mathbf{R}_0 = initial radius vector (inserted).

This basic inertial navigation equation points up some of the basic characteristics of inertial systems:

(1) The inertial system must have initial position and initial velocity information (the two constants of integration).

(2) The accelerometer senses all non-gravitational forces (including thrust, drag, lift, and structural support).

(3) The gravitational field is not sensed; it must be calculated from known field equations.

Performance Characteristics. Inertial systems have several distinct characteristics not common to other systems. They

(1) give continuous rather than discrete information on acceleration, velocity, position, and vehicle attitude;

(2) require no signals from outside the system, so they are jamproof and can be used in vehicles launched in salvo;

(3) do not radiate signals, so they are difficult to detect in military applications;

(4) can be launched quickly but are most accurate when adequate prelaunch time is available for warmup, trim, and alignment;

(5) have errors that are a function of time rather than speed or distance;

(6) can provide by-product signals such as stabilization for flight control or radar antennas, velocity for mapping cameras, etc.

Error Characteristics. Systematic errors in pure inertial systems based on error in the knowledge of the gravity vector have a characteristic

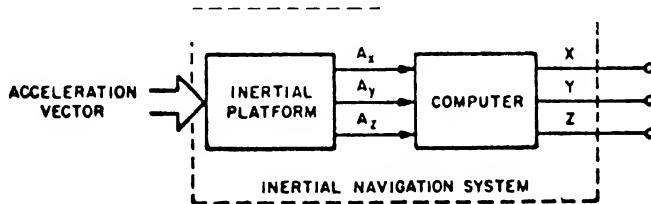


FIG. 1. Function of the basic inertial system.

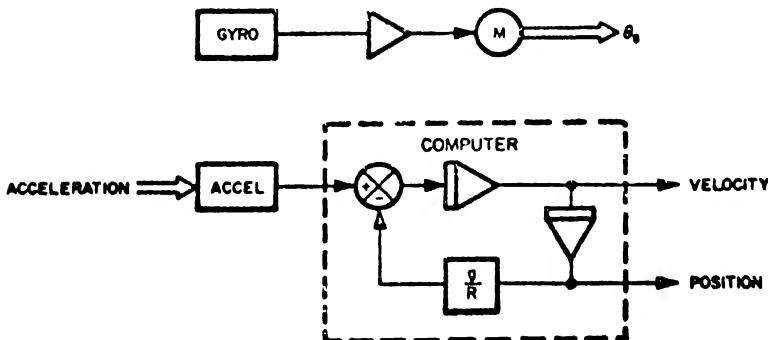


FIG. 2. Space-stabilized system block diagram.

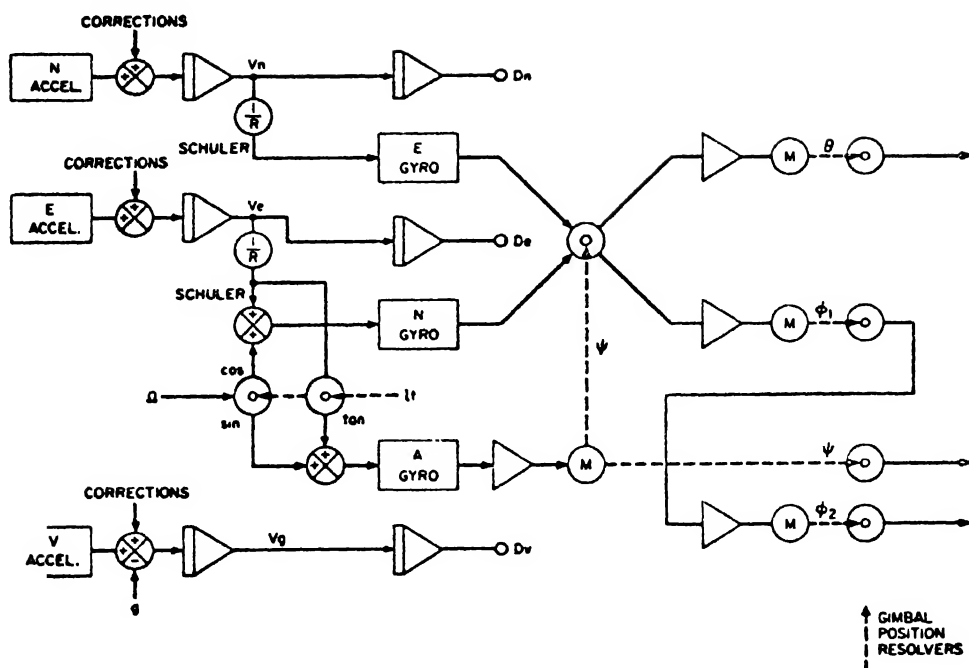


FIG. 3. Local vertical system block diagram.

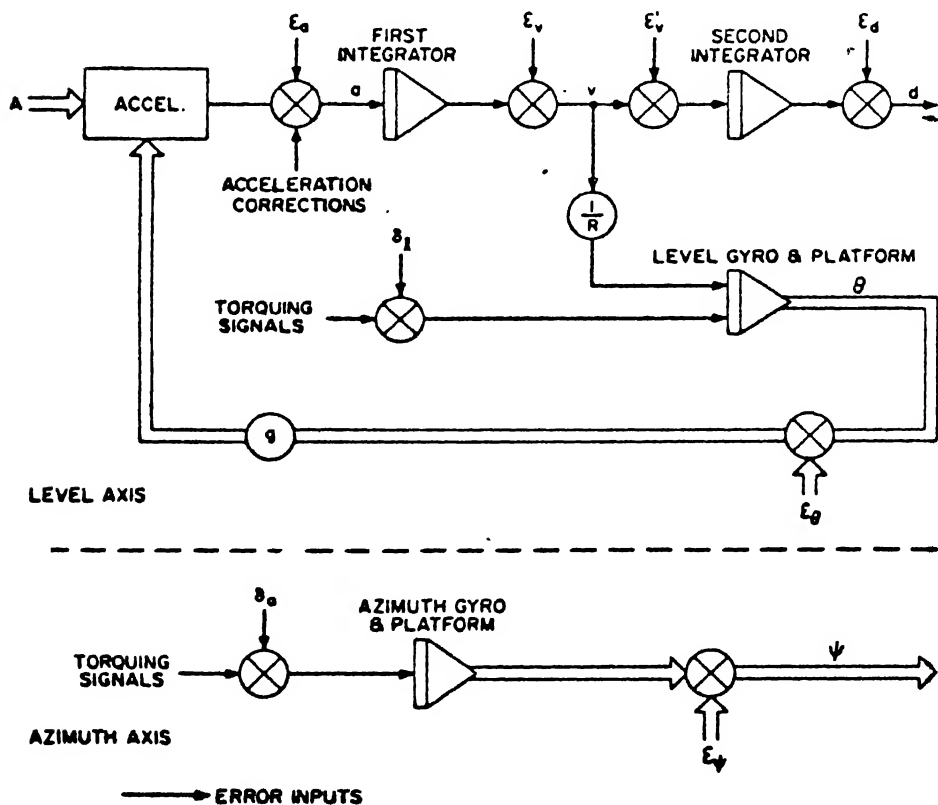


FIG. 4. Error inputs to local vertical system.

Schuler oscillation corresponding to orbital period (84.5 minutes at the earth's surface) in the horizontal components of the navigation position vector and an unstable exponential error in the local vertical component.

Basic System Mechanization. In the basic space-stabilized system (Fig. 2), the three accelerometer input axes are stabilized to any desired orientation in space by the GYROSCOPE (q.v.) stabilization control loops. A popular orientation for space vehicle launches is to have the Z axis vertical at time $t = 0$, the Z and X axes in the orbital or launch plane, and the Y axis orthogonal to Z and X . The gravitational force which starts out parallel to the Z axis is continuously computed as a function of the vehicle's position.

Local Vertical System. Another commonly used system mechanization is the local vertical system that maintains the Z axis vertical and the X axis north throughout flight for convenience in surface and air navigation. This requires biasing each of the gyros with a turning rate which

is a function of the earth's rotation rate plus the vehicle angular velocity around the curved earth's surface:

$$\omega_x = \Omega \cos lt + \frac{V_e}{R}$$

$$\omega_y = \frac{V_n}{R}$$

$$\omega_z = \Omega \sin lt + \frac{V_e}{R} \tan lt$$

where

- $\omega_{x,y,z}$ the computed bias signals to the x, y, z (north, east, and vertical) gyros
 Ω earth's sidereal rotation rate (15.041 deg/hr)
 lt local latitude of the vehicle
 $V_{e,n}$ vehicle east and north velocity
 R local earth's radius plus altitude.

In the local vertical system the accelerometers measure acceleration in a north, east, and vertical

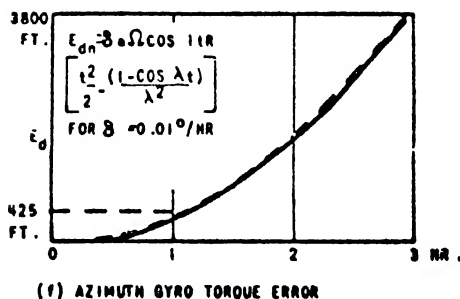
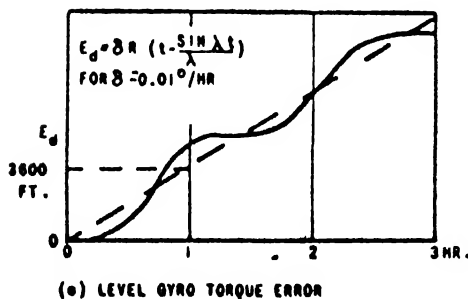
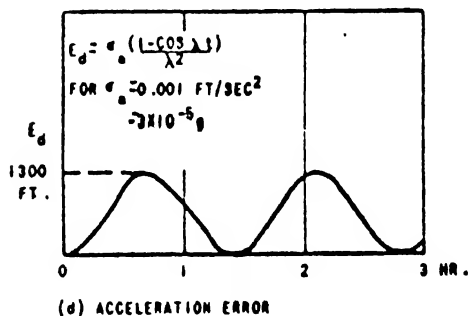
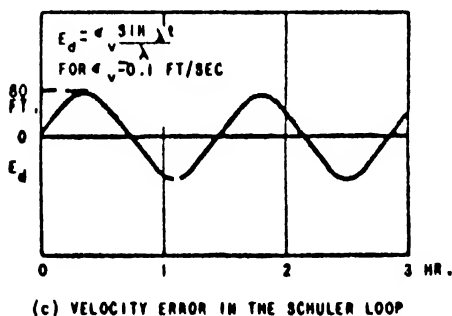
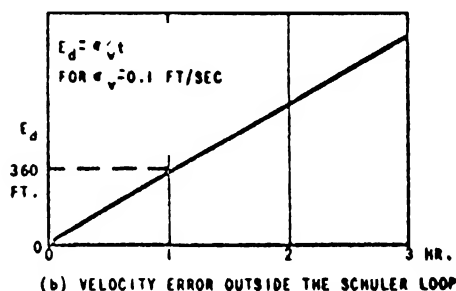
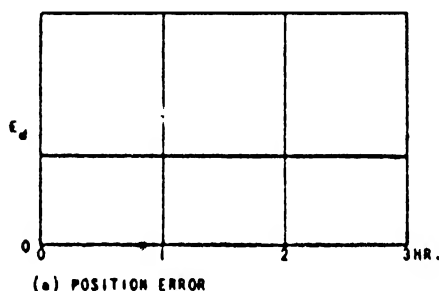


FIG. 5. Classes of position errors.

reference system which rotates in space and therefore requires CORIOLIS (q.v.) corrections. The explicit gravitational calculation is avoided in the two level axes, and the vertical axis is often unnecessary in two-dimensional surface navigation. This mechanization is shown in Fig. 3.

Errors in Local Vertical Systems. The characteristic errors can be seen by analyzing the dynamic effect of errors entering the system at various points (Fig. 4). The position errors resulting from error inputs of assumed magnitude in each case are shown in Fig. 5. The total error in calculated position will be some combination of these.

Self-alignment. During the navigation mode the gyros hold the stable element to the desired attitude, but this attitude must be assumed before the navigation mode begins. This can be done during a self-alignment mode. The stable element is placed in a local level attitude by using information from the two level accelerometers, whose outputs are null when their input axes are level. Azimuth orientation information is derived from the east gyro, which senses no component of the earth's rate of rotation when oriented east. The self-level and gyrocompass alignment mode is shown in Fig. 6.

Strapdown Systems. In addition to the gimballed gyro-stabilized systems, there are several gimballess or strapdown configurations. One is the

relatively simple accelerometer, vehicle attitude control system first described. Another uses three accelerometers whose vehicle-referenced output pass through a dynamic coordinate transformation matrix in the computer. Three *rate* gyros provide the computer with vehicle attitude rate information necessary to compute the matrix. Although this approach eliminates gimbaling, it requires precision rate gyros and relatively large computer capacity to integrate attitude rate and provide a dynamic matrix.

A third gimballess system concept uses an inertial reference such as electrostatically suspended gyros to give vehicle *attitude* information to the computer for use in calculating the dynamic coordinate conversion matrix. Attitude rate integration is thereby avoided.

Steering Commands. An inertial navigation system primarily provides vehicle velocity or position. Guidance can be provided by the inertial system by supplying target location information to the computer, which then compares vehicle position with target position and calculates steering and (in the case of space flight) engine shutoff commands.

RICHARD H. PARVIN

References

- Parvin, Richard H., "How Coriolis Works," *Electronic Industries*, April, 1960.

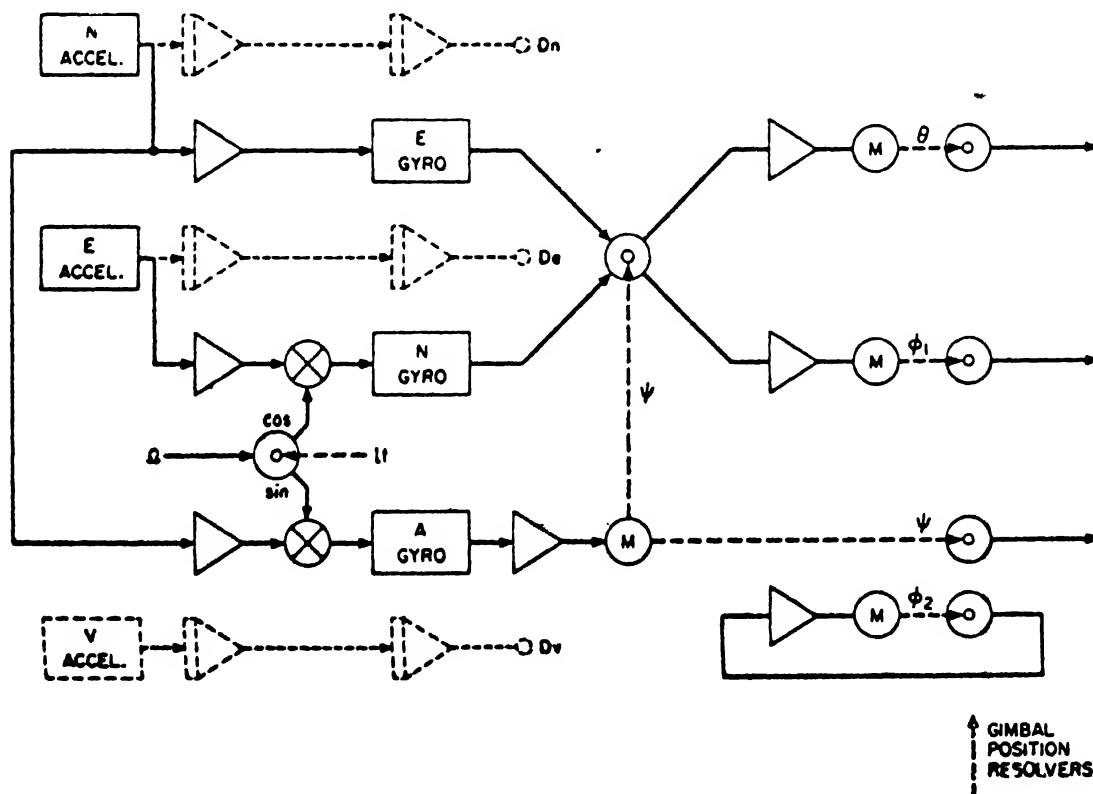


FIG. 6. Self-alignment mode.

Stearns, Edward V., "Navigation and Guidance in Space," Englewood Cliffs, N.J., Prentice-Hall, 1962.

Pitman, George R., "Inertial Guidance," New York, John Wiley & Sons, 1962.

Parvin, Richard H., "Inertial Navigation," Princeton, D. Van Nostrand Co., 1962.

Cross-references: ASTRODYNAMICS, ASTRONAUTICS, GYROSCOPE.

INFRARED RADIATION

The region of the electromagnetic spectrum between the wavelength limits 0.7 and 1000μ (7×10^{-7} and 1×10^{-1} cm) has become known as infrared radiation. The lower wavelength limit is set to coincide with the upper limit of the visible radiation region. Radiation of wavelength greater than 1000μ is generally thought of as the microwave spectrum. Both limits are arbitrary, and represent no change in characteristics as they are passed. Conventionally, the region between 0.7 and 1.5μ is called the *near infrared region*; that between 1.5 and 20μ , the *intermediate infrared region*; and that between 20 and 1000μ , the *far infrared region*.

For many applications, the location of infrared radiation in the spectrum is described by its wavelength in microns, μ , ($1\mu = 10^{-4}$ cm). In applications where the relative energy of the radiation is of interest, the *wave number*, σ , is used. The wave number is defined as the reciprocal of the wavelength, λ , in centimeters, and is expressed in units of cm^{-1} (called the *kayser*). This quantity is used more commonly than the frequency ν of the radiation, which is related to σ as follows:

$$\sigma = \nu/c \quad (1)$$

where c is the velocity of light.

Infrared radiation is produced principally by the emission of solid and liquid materials as a result of thermal excitation and by the emission of molecules of gases. Thermal emission from solids is contained in a continuous spectrum, whose wavelength distribution is described by the relation

$$I_\lambda d\lambda = \frac{2\pi c^2 h \epsilon_\lambda}{\lambda^5} \frac{1}{e^{hc/\lambda k T} - 1} d\lambda \quad (2)$$

where

I_λ = spectral radiant emittance of the solid into a hemisphere in the wavelength range from λ to $(\lambda + d\lambda)$

h = Planck's constant 6.62×10^{-27} erg sec

ϵ_λ = spectral emissivity

k = Boltzmann's constant $= 1.38 \times 10^{-16}$ erg $^\circ\text{K}$

T = absolute temperature of the solid emitter, $^\circ\text{K}$.

The spectral emissivity, ϵ_λ , is defined as the ratio of the emission at wavelength λ of the object to that of an ideal blackbody at the same temperature and wavelength. When ϵ_λ is unity, Eq. (2) becomes the Planck radiation equation for a blackbody.

The distribution of radiant emittance with wavelength for blackbody radiators at different temperatures is shown in Fig. 1. It is apparent from the Figure that blackbody radiation from emitters at temperatures below about 2000°K falls predominantly in the infrared region. An emitter which exhibits a constant value less than unity of spectral emissivity at all wavelengths is called a *gray-body* radiator. Most solid radiators show a general decrease in spectral emissivity with increasing wavelength in the infrared; however, over limited spectral ranges, many materials are approximately gray-body radiators. Radiators which approach the characteristics of ideal blackbodies can be made in the form of uniformly heated cavities. A relatively small aperture, through which the cavity can be observed, serves as the source of blackbody radiation.

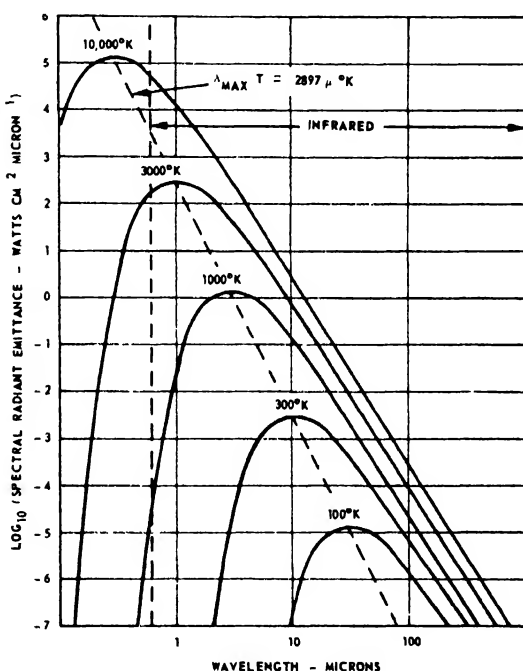


FIG. 1. Spectral radiant emittance of a blackbody at various temperatures.

Infrared radiation is also observed as emitted from excited molecules of gases. Many of the energy transitions which take place in gases excited thermally or electrically result in radiation emission in the infrared region. Gaseous emission differs in character from solid emission in that the former consists of discrete spectrum lines or bands, with significant discontinuities, while the latter shows a continuous distribution of energy throughout the spectrum. The predominant source of molecular radiation in the infrared is the result of vibration of the molecules in characteristic modes. Energy transitions between various states of molecular rotation also produce infrared radiation. Complex molecular gases radiate intricate spectra, which may be analyzed to give

information of the nature of the molecules or of the composition of the gas.

The propagation of infrared radiation through various media is, in general, subject to absorption which varies with the wavelength of the radiation. Molecular vibration and rotation in gases, which are related to the emission of radiation, are also responsible for resonance absorption of energy. The gases in the atmosphere, for example, exhibit pronounced absorption throughout the infrared spectrum. The principal gases of the atmosphere, nitrogen and oxygen, do not absorb significantly in the infrared region. However, the lesser constituents, water vapor (H_2O), carbon dioxide (CO_2), and ozone (O_3), are responsible for strong absorption in the infrared. The absorption of radiation is so prevalent that those spectral bands in which relatively little absorption occurs are identified as *atmospheric windows*.

Solid and liquid materials show, as a rule, strong absorption in the infrared. There are, however, many solids which transmit well in broad regions of the infrared spectrum. Many materials, such as water and silica glasses, which show little absorption in the visible, are opaque to infrared radiation at wavelengths greater than a few microns. Many of the electrically insulating crystals, such as the alkali halides and the alkaline-earth halides, which transmit well in the visible, also are transparent to much of the near and intermediate infrared spectrum. Several of the semiconductor materials absorb strongly in the visible, but become transparent in the infrared beyond certain wavelengths characteristic of the semiconductor.

Detection of the presence, distribution and or quantity of infrared radiation requires techniques which are, in part, unique to this spectral region. The frequency of the radiation is such that essentially optical methods may be used to collect, direct, and filter the radiation. Transmitting optical elements, including lenses and windows, must be made of suitable materials, which may or

may not be transparent in the visible spectrum. Table I gives characteristics of several transmitting materials suitable for use in infrared optical systems. To avoid chromatic aberration, reflecting mirrors are commonly used in infrared optical systems to focus and deviate the radiation when broad spectral bands are observed. Filters for the infrared are designed and constructed like those for the visible, except for the choice of materials and, in the case of interference filters, the thickness of the layers.

The detector element for infrared represents the most unique component of the detection system. Photographic techniques can be used for the near infrared out to about 1.3μ . Photoemissive devices, comparable to the visible- and ultraviolet-sensitive photocells, are available with sensitivity also extending to about 1.3μ . The intermediate infrared region is most effectively detected by photoconductors. These elements, photosensitive semiconductors, are essentially photon detectors, which respond in proportion to the number of infrared photons in the spectral region of wavelength. This wavelength corresponds to the minimum photon energy necessary to overcome the forbidden gap of the semiconductor. A number of sensitive photoconductors are available with spectral cutoff at various wavelengths in the infrared. Photoconductors are employed as resistive elements, as photovoltaic $p-n$ diodes, or as photoelectromagnetic elements, according to the particular electrical advantage to be gained. All spectral regions from ultraviolet through visible, infrared, and microwaves, can be detected by an appropriately designed thermal element, which responds by being heated by the absorption of the incident radiation. In the infrared, thermal detectors take the forms of thermocouples, bolometers, and pneumatic devices. The thermal elements, in general, are not as sensitive or as rapidly responding as photoconductors in spectral regions where they both respond. However, the broad spectral response and uniform energy

TABLE I. INFRARED TRANSMITTING MATERIALS

Material	Useful Transmission Region (μ)	Refractive Index Near Transmission Peak	Special Characteristics
Optical glasses	0.3-2.7	1.48-1.70	Best for near infrared
Fused silica	0.2-3.5-4.5	1.43	Some types show absorption near 2.7μ
Arsenic trisulfide	0.6-12.0	2.4	A glass; subject to striations
Calcium aluminate	0.3-5.5	1.8	A glass; subject to attack by water
Sapphire	0.17-6.0	1.7	Single crystal, hard, refractory
Silicon	1.1-20	3.4	Low density; opaque to visible
Germanium	1.8-20	4.0	Opaque to visible
NaCl	0.2-15	1.52	Water soluble
KBr	0.21-27	1.54	Water soluble
LiF	0.11-6	1.35	Low solubility in water
CaF ₂	0.13-9	1.41	Insoluble
Thallium bromide-iodide (KRS-5)	0.5-40	2.38	Fairly soft; cold flows
AgCl	0.4-25	2.0	Soft; cold flows

sensitivity characteristics make them highly useful. Table 2 gives representative characteristics of several commonly used infrared detectors.

The most common application of infrared radiation is, of course, radiant heating. Solid radiators, such as hot tungsten filaments, alloy wires, and silicon carbide rods are employed extensively as sources of infrared to provide surface heating by radiation.

ployed to eliminate the effects of the short-wavelength radiation.

Optical-electronic devices of many varieties have been designed to determine the direction of weakly radiating remote objects by means of detection of their infrared emission. Military applications have been found which have been made possible uniquely by this technique. Missiles can be guided to their target by infrared detection

TABLE 2. INFRARED DETECTORS

Detector (operating temperature)	Region (μ)	Relative Detectivity at Peak	Time Constant (sec)	Special Features
PbS (295 K)	Visible-2.8	1	2×10^{-4}	Thin-film photoconductor
PbSe (195 K)	Visible 5.6	.06	2×10^{-4}	Thin-film photoconductor
InSb (77 K)	1-5.6	.5	2×10^{-7}	Photovoltaic crystal
Ge (Hg doped) (25 K)	1-16	8	10^{-6}	Photoconductor crystal
Ge (Cu doped) (5 K)	1-29	.5	10^{-8}	Photoconductor crystal
Ge (Zn doped) (5 K)	1-40	2	10^{-8}	Photoconductor crystal
Thermistor bolometer (295 K)	All	2.8×10^{-3}	$10^{-3} - 10^{-2}$	Flake of mixed oxides
Golay cell (255 K)	All	.02	1.5×10^{-2}	Pneumatic
Thermocouple (295 K)	All	.002	1.5×10^{-2}	Used in spectrometers

Infrared spectroscopy has become a powerful analytical tool in the chemistry laboratory. Organic molecules, in general, contain interatomic valence bonds which exhibit characteristic resonance frequencies which can be identified in the absorption spectrum of the material in gaseous form. Such information can be used to study the structure of complex molecules. It also serves in aiding the identification of the presence of known valence bonds in chemical analysis. Most absorption lines and bonds due to molecular vibrations fall in the frequency range 500 to 5000 cm^{-1} (wavelength range 2 to 20 μ). A large quantity of data has been gathered on the detailed absorption spectra of many gaseous materials. The characteristic spectra of many organic molecules are such that identification of the presence of the molecules, as well as the presence of particular radicals within the molecules, can be readily observed. Petroleum chemistry, for example, has been greatly aided by the application of infrared spectroscopy to the identification of many of the complex constituents in petroleum products.

Observation of infrared absorption spectra is carried out by means of an infrared spectrophotometer, in which the transmission of monochromatic radiation by a gaseous sample in a cell is compared with that of a blank cell, while the wavelength of the radiation is scanned through the spectral range of interest. Prism dispersing elements are usually used in the infrared, rather than gratings, because of the difficulty with the latter of separating the several orders in the wide spectral range covered. Far infrared spectroscopy is complicated by the omnipresence of background and scattered radiation of shorter wavelength emitted inside the instrument at room temperature. Special techniques of filtering must be em-

ployed to eliminate the effects of the short-wavelength radiation.

Such devices require the detection of low-level radiation in the intermediate infrared region. Optical lenses or mirrors are used to collect the observed radiation and concentrate it onto the sensitive infrared detector. High-gain, low-noise electronic amplifiers must be provided to increase the weak signal from the detector to a level which can be used to operate controls or displays, as demanded by the application. Optical filtering is applied in order to restrict the observed spectral region to one in which the target is effectively detected, with a minimum of interference from radiation from its background. The wavelength of detection is such that angular resolution capability, as set by diffraction, is much greater with infrared devices than that of radar devices. Detection of targets at great distances through intervening atmosphere is more effective in the infrared than in the visible because of the much lower atmospheric scattering in the infrared.

Detailed discussions of the characteristics, detection and applications of infrared radiation may be found in the references.

R. H. McFEE

References

- Jamieson, J. A., McFee, R. H., Plass, G. N., Grube, R. H., and Richards, R. G., "Infrared Physics and Engineering," New York, McGraw-Hill Book Co., 1963.

Smith, R. A., Jones, T. E., and Chasmar, R. P., "The Detection and Measurement of Infrared Radiation," Fair Lawn, N.J., Oxford University Press, 1957.

Herzberg, G., "Infrared and Raman Spectra of Polyatomic Molecules," Princeton, N.J., D. Van Nostrand Company, 1945.

Szymanski, H. A., and Alpert, N. A., "IR: Theory and Practice of Infrared Spectroscopy," Plenum Press, 1964.

Kruse, P. W., McGlauchlin, L. D., and McQuistan, R. B., "Elements of Infrared Technology," New York, John Wiley & Sons, 1962.

Cross-references: ABSORPTION SPECTRA; LIGHT; RADIATION; THERMAL; SPECTROSCOPY.

INTERFERENCE AND INTERFEROMETRY

Interference is the term used to signify a large class of phenomena in light, and interferometry is the technique of high-precision measurement based on these phenomena. Ordinarily rays of light crossing the same point in different directions do not interfere with each other; each ray is propagated as though it alone were present. However, there are certain interesting cases when these rays do interfere with each other; the interference may be destructive, as when they cancel each other's effect, or may be constructive, as when they reinforce each other. Interference is a consequence of light being propagated in the form of waves.

Young's Experiment. The classic experiment in interference is the one performed by Thomas Young in 1802. A source of light SL (see Fig. 1)

ripples be produced by a vibrating metal strip dipping in and out of the surface of water. A light floating body, say a leaflet, wobbles up and down with the same frequency as the vibrator. The ripples spread out in widening circles. If now another vibrator of the same frequency is brought near the first, the appearance of the ripples is completely changed. Along certain radial lines starting from the two vibrators, the water surface seems undisturbed; the leaflet does not move if placed along these lines. In between these lines the ripples have a very large amplitude.

At points equidistant from the two vibrators, i.e., along the perpendicular bisector of the line joining the vibrators, the waves from both arrive in phase. Both systems of waves tend to move the water up or down at the same time. The amplitude of the waves along this line is double that due to either of the systems of waves. At some other point sufficiently away from the perpendicular bisector, the crest of one system arrives at the same time as the trough of another, and thus the two systems cancel each other. Destructive interference occurs if the distances of the two vibrators from the point differ by $(n + \frac{1}{2})\lambda$, where λ is the wavelength and n is zero or an integer. Constructive interference occurs if the path difference is $n\lambda$.

An important condition for interference is that the two systems of waves are coherent, i.e., that they always have the same phase relation to each other. If the two slits P and Q in Young's experiment were illuminated by waves from two different sources, interference effects would not be observed. The reason is that waves produced by two sources would have no phase



FIG. 1. Young's experiment.

that is placed behind a narrow slit illuminates two other slits P and Q which are parallel and very close to each other. At some distance away is a screen which receives the light from the two slits. On the screen is seen a series of bright and dark fringes. If either of the two slits is covered, the fringes disappear, and the screen is almost uniformly illuminated. The combined effect due to the two slits is that at certain points there is no light at all, and at other points the brightness is four times that due to a single slit.

This puzzling phenomenon of light upon light producing darkness can be readily understood by considering the analogous case of ripples on the surface of water. Let a continuous series of

relation to each other. When the two slits are illuminated by different parts of the same wave front, they always arrive at any point beyond the slit with a constant difference in phase.

Theory. The mathematical expression for a progressive wave is

$$S = a \cos 2\pi\nu \left\{ \left(t - \frac{x}{v} \right) + \alpha \right\} \quad (1)$$

where S is the magnitude of the electric or magnetic field, also called displacement, at time t and distance x , a is the amplitude, ν the frequency, v the velocity, and α a term denoting

the phase. Due to the superposition of two waves, denoted by subscripts 1 and 2,

$$S = S_1 + S_2 = a_1 \cos 2\pi\nu \left\{ \left(t - \frac{x}{v} \right) + \alpha_1 \right\} + a_2 \cos 2\pi\nu \left\{ \left(t - \frac{x}{v} \right) + \alpha_2 \right\} \quad (2)$$

Both waves have the same frequency and velocity but different amplitudes and phases; x is measured from any arbitrary point. The displacement at $x = 0$ is given by

$$S = A \cos (2\pi\nu t + \alpha) \quad (3)$$

where A is the amplitude and α the phase of the resulting wave. By expanding the right-hand sides of Eq. (2) with $x = 0$ and Eq. (3), equating coefficients of $\cos 2\pi\nu t$ and $\sin 2\pi\nu t$, squaring and adding, one can see that

$$A^2 = a_1^2 + a_2^2 + 2a_1a_2 \cos(\alpha_1 - \alpha_2) \quad (4)$$

As $\cos(\alpha_1 - \alpha_2)$ varies between $+1$ and -1 , A varies between $(a_1 + a_2)$ and $(a_1 - a_2)$. If the amplitudes a_1 and a_2 are both equal to a , the minimum value of A is 0 and the maximum value is $2a$. Since the intensity is proportional to the square of the amplitude, the intensity at the maximum is 4 times that due to either wave. The condition for maximum brightness is: $\alpha_1 - \alpha_2 = 2n\pi$, which corresponds to a path difference of $n\lambda$.

Let R and S, Fig. 1, be the positions of the central fringe and the n th bright fringe below R. Let PQ = s , RS = x , and let D be the distance of the screen from the two slits.

$$PS^2 - QS^2 = \left\{ D^2 + \left(x + \frac{s}{2} \right)^2 \right\} - \left\{ D^2 + \left(x - \frac{s}{2} \right)^2 \right\} = 2xs$$

$$PS - QS = 2xs / (PS + QS) \approx 2xs / 2D$$

since $PS + QS$ is very nearly equal to $2D$.

$$PS - QS = n\lambda, \text{ so that } \lambda = \frac{1}{n} \frac{xs}{D} \quad (5)$$

Equation (5) formed the basis for the first experimental determination of the wavelength of light.

Young's experiment in its original form was difficult to perform and failed to carry conviction when the results were first published. If $s = 1$ mm, $D = 2$ meters, the distance between successive

fringes for sodium yellow light ($\lambda = 5.89 \times 10^{-5}$ cm) is only 1.2 mm. The illumination is too poor, the fringes are too close, and two fine slits at 1 mm distance are difficult to produce. The controversy as to whether light is propagated as corpuscles or waves had existed for over a century and a half. Francesco M. Grimaldi, who is regarded as the founder of the wave theory of light, in his book, *Physico-Mathesis de Lumine, Coloribus et Iride*, published in 1665, described several experiments on diffraction and interference of light, and presented the rudiments of a wave theory. Newton discussed several diffraction effects in his *Opticks*, published in 1708; he threw his weight heavily on the side of the corpuscular theory of light. Experiments more convincing than those of Young were needed to overthrow a theory based on Newton's authority. Between 1814 and 1816, Fresnel introduced two better methods of producing interference fringes, he also gave a more complete theory of the formation of the fringes, based on the hypothesis of secondary wavelets which was first developed in 1678 by Huygens. Huygens' hypothesis was that every point on a wave front acted as the source for a secondary train of waves and that the envelope of these secondary waves determined every successive position of the wave front.

The two improved experimental arrangements introduced by Fresnel were the bimirror and the biprism. These solve the difficulty of obtaining two slits sufficiently narrow and close to each other. In the bimirror arrangement, light from a narrow slit is reflected by two plane mirrors inclined at a small angle to each other. Thus the two slits are replaced by the two images of a single slit, and the distance between these can be adjusted by changing the angle between the mirrors. In the biprism arrangement, two small angle prisms joined at their base each produce a small deflection of the light emerging from a single slit, and thus cause two sets of coherent waves to be superposed. A single mirror may also be used as devised by Lloyd to produce interference between wave trains produced by a slit and its image.

Applications of interference effects. There are many interesting applications of the principle of interference. Refractometers based on interference effects are used to measure small changes in the refractive index of transparent media (see REFRACTION). In the Rayleigh refractometer (see Fig. 2), light from a linear source S, made parallel by a lens L_1 , is split into two beams by two fairly wide slits, and then made to pass through two similar tubes T_1 and T_2 . After transmission

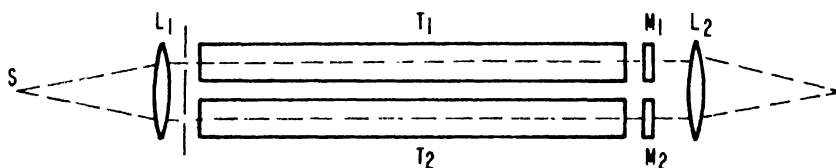


FIG. 2. Rayleigh refractometer.

through the tubes, the two beams are brought to a common focus by another lens L_2 . If the two tubes contain transparent media of the same refractive index, say the same liquid, the center of the fringe pattern is formed on the axis of the instrument. If the refractive index of the liquid in one of the two tubes is changed, as for example by introducing a solvent, the fringes shift across the focal plane of the viewing lens. By counting the number of fringes which cross a reference line, the equivalent path difference and hence the change in refractive index can be calculated. Compensator plates M_1 and M_2 , which restore the fringe pattern to its original position, are convenient devices for counting the fringe shift.

Michelson's method of measuring stellar diameters is another application of interference. A beam mounted over the entrance aperture of a large telescope carries four mirrors as shown in Fig. 3. The arrangement is similar to that of

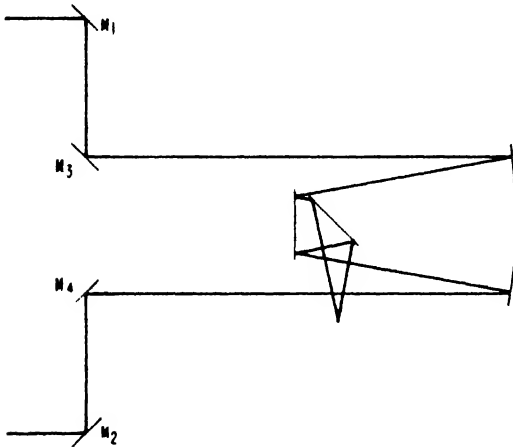


FIG. 3. Stellar interferometer.

Young's experiment, with the mirrors M_1 and M_2 as the slits and the star as the source. The diffraction image of the star is crossed by interference bands if M_1 and M_2 are relatively close to each other. As the distance between them is increased, the fringes become less distinct and finally disappear. This is due to the fringe pattern from one half of the star being completely canceled by that from the other. If the distance between M_1 and M_2 for disappearance of the fringes is s , r is the distance of the star, d its diameter, and λ the wavelength of light; $d/r = 1.22\lambda/s$. The first star to be measured by this method was Betelgeuse, for which the bands disappeared at $s = 306.5$ cm. Substituting values 5.75×10^{-6} cm for λ , and 1.712×10^{20} cm for r , as determined from the parallax of Betelgeuse, the diameter of the star is found to be 3.918×10^{13} cm, which is 31 per cent greater than the diameter of the earth's orbit around the sun. Several other near stars of large size have since been measured by the same method.

Thin films of transparent media produce striking interference effects, as for example,

when a few drops of gasoline are spilled over a wet pavement. The two wave trains which interfere in this case are those reflected from the upper surface of the oil film and from the water-oil interface. The colors of butterflies' wings and sea shells have a similar origin. The so-called Newton's rings are produced by the air film between two partially reflecting, spherical surfaces. The optical quality of a glass surface can be tested by causing interference fringes between it and a standard test plate of high optical quality. The fringes are analogous to contour lines in geographical maps, each new fringe indicating deviation from true flatness by half a wavelength.

Interferometers. Interferometers are instruments of high-precision measurement based on the principle of interference. A beam of light is divided into two or more beams by partial reflection and transmission, and these are recombined after they have traveled different path distances. Of the many different types of interferometers, only two which are widely used will be described here, the Michelson interferometer and the Fabry-Perot interferometer.

The Michelson interferometer is shown schematically in Fig. 4. Monochromatic light from an

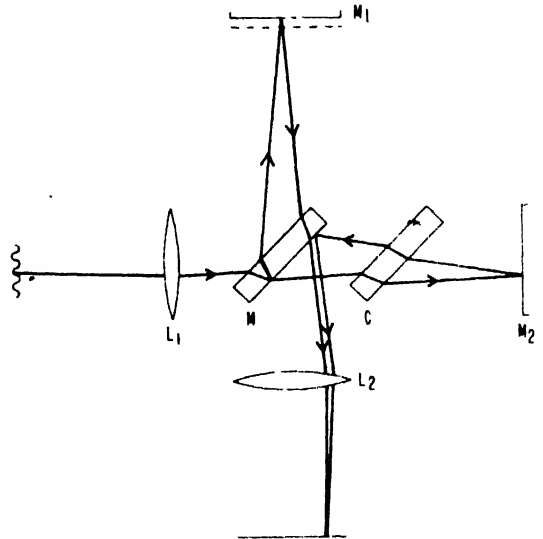


FIG. 4 Michelson interferometer.

extended source is collimated by a lens L_1 and falls on the beam splitter M of which the hind surface partially reflects half the intensity upward and transmits the other half. The two halves of the beam are returned by the mirrors M_1 and M_2 , and the interference pattern is viewed in the focal plane of the lens L_2 . C is a compensator plate similar to M , which gives the beam from M_2 an extra path equal to that traveled by the other beam in the beam splitter. The form of the fringes depends on the adjustment of the two mirrors M_1 and M_2 . If they are not quite at right angles to each other, and the difference in path of the two beams is small, the image of M_2 in M forms

a thin wedge with the front surface of M_1 . The fringes are straight and parallel to the apex of the wedge. If the difference in path is large, the two mirrors should be adjusted so that the image of M_2 in M is exactly parallel to M_1 . In this case, the fringes are circular. Each circle is due to pencils of light which have a constant inclination to the axis of the lens L_1 .

Two of the applications of the Michelson interferometer are of historic importance: the standardization of the meter in terms of the wavelength of light and the Michelson-Morley experiment for the drift of ether. If one of the two mirrors is moved parallel to itself, the pattern of fringes shifts across the field of view. The displacement of the mirror for each fringe shift is $\lambda/2$. This method was first used by Michelson and Benoit in 1892 for comparing the red line (6438Å) of cadmium with the International Prototype Meter which is kept in Paris, France. More precise measurements of the meter in terms of the wavelength of light have since been

moved parallel to itself at a very constant rate. The variation of intensity over a small area at the center of the ring system at P is measured by an infrared detector. If the source were strictly monochromatic, the output signal of the detector would vary sinusoidally with time. The displacement of the mirror between successive maxima is half a wavelength. With a composite source as input, the output is the sum of a large number of sine functions, each of them being due to the energy in a narrow wavelength band of the source. A Fourier transform of the output signal, as may well be obtained with the aid of a digital computer, gives the spectral energy distribution of the source. The compactness of the instrument is a special advantage compared to infrared prism monochromators, and hence several designs of the interferometer spectrophotometer have been developed for use in satellites and space probes.

The Fabry-Perot interferometer (see Fig. 5) consists of two parallel plates of glass or quartz.

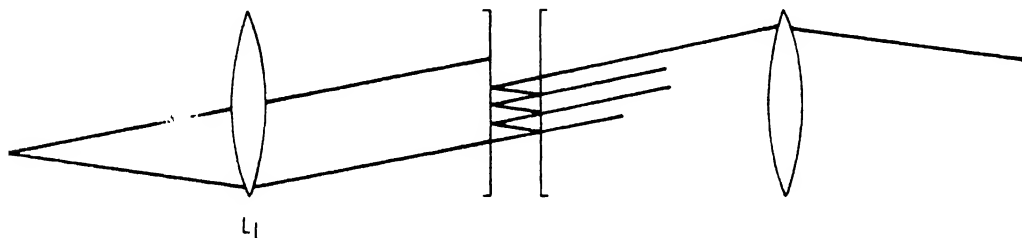


FIG. 5 Fabry-Perot interferometer.

made by other observers. Since the wavelength of light is a more reliable standard and can be measured with greater accuracy, it was judged desirable to replace the standard meter by a suitable spectral line as the standard of length. The International Commission of Weights and Measures formally adopted in 1960 the orange-red line of krypton 86 as the standard and defined the meter as exactly 1 650 763.73 wavelengths in vacuum of this line.

The MICHELSON-MORLEY EXPERIMENT of 1887 was an attempt to measure the speed of the ether "wind" past the moving earth. If one arm of the interferometer is in the direction of the earth's motion relative to the ether and the other at right angles to this motion, the relative path difference between the two beams of light is nearly Lv^2/c^2 , where L is the length of each arm, v is the velocity of the earth and c is the velocity of light. By floating the interferometer in a pool of mercury and rotating it through 90°, a fringe shift corresponding to twice this path difference should be observed. Accurate experiments showed that the fringes did not shift. This negative result served as the basis for the theory of relativity.

A recent application of the Michelson interferometer is for spectrophotometry of composite sources. The method is especially applicable for the infrared range. One of the two mirrors is

The inner surfaces are optically flat and semi-silvered. Light from an extended source is made parallel by a lens L_1 , passes through the interferometer and is focused by another lens L_2 . A system of circular fringes is observed in the focal plane of L_2 . The path of an oblique ray of light in between the two plates is shown in Fig. 5. Interference takes place between the directly transmitted ray and the rays that undergo one or more reflections between the plates. A high degree of wavelength resolution is the main advantage of the Fabry-Perot interferometer. In a typical case of 1-cm separation of plates and 90 per cent reflectance, wavelength resolution is over a million; i.e., two wavelengths 0.005Å apart at 5000Å will give completely distinguishable ring systems. By increasing the reflectance of the plates or the separation between them, the wavelength resolution can be increased to any desired degree. Laser beams which give highly coherent single wavelengths permit plate separation of over a meter. Extensive use has been made of the Fabry-Perot interferometer for precision measurement of wavelengths of spectral lines.

M. P. THEKAKARA

References

- Born, M., and Wolf, E., "Principles of Optics," Chs. 7, 10 and 11, New York, Pergamon Press, 1959.

Strong, J., "Concepts of Classical Optics," Chs. 8, 11, 12 and Appendix F, San Francisco, W. H. Freeman and Co., 1958.

Mollet, P., Ed., "Optics in Metrology," Colloquia of the International Commission for Optics, New York, Pergamon Press, 1960. Several excellent articles on recent applications.

Jenkins, F. A., and White, H. A., "Fundamentals of Optics," Chs. 12, 13, 14 and 17, New York, McGraw-Hill Book Co., 1957.

Andrews, C. L., "Optics of the Electromagnetic Spectrum," Chs. 6, 7 and 18, Eaglewood Cliffs, N.J., Prentice-Hall, Inc., 1961.

Cross-references: DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS, ELECTROMAGNETIC THEORY, LIGHT, MICHELSON-MORLEY EXPERIMENT, REFRACTION, RELATIVITY, WAVE MOTION.

INTERNATIONAL GEOPHYSICAL YEAR AND INTERNATIONAL YEARS OF THE QUIET SUN

It has been recognized for some time that the sun, ultimate source of practically all of the energy utilized on earth (the only current exception being the relatively small amount of power produced by nuclear fuels), influences many earthly phenomena. Aside from the obvious solar control of the weather through visible and invisible light—electromagnetic radiation—continuously emitted by the sun, other less well-understood effects occur. For example, the so-called earth storm (not to be confused with a weather disturbance near the ground) results from the arrival not only of radiation but also of matter—streams of electrically charged particles—spewed out by the sun following violent eruptions or flares in the chromosphere, just above the sun's "surface" (photosphere).

A multitude of geophysical effects arises from the interactions of these radiation and particle fluxes with the upper reaches of the earth's atmosphere, manifesting themselves as phenomena such as magnetic storms, radio disturbances, and auroral displays. The frequency of occurrence of these transitory happenings in the upper atmosphere waxes and wanes as the level of solar activity changes during the well-known sunspot cycle.

Taken together, the International Geophysical Year (IGY) and the International Years of the Quiet Sun (IQSY) mark the beginning and end points of a tremendously significant study of the extremes of the solar cycle. The first part, IGY (1957–1958), took place at the peak of sunspot activity (it was not only peak for the average eleven-year cycle—but, in intensity, has probably not been matched at least since Galileo first observed sunspots in the early seventeenth century). IQSY (1964–1965) is the follow-up aimed at catching the myriad interrelated phenomena of sun, space, and earth at the sun's nadir of activity. But the current enterprise is not to be regarded as a small-scale repetition of its illustrious predecessor. On the contrary, in the fields

that it embraces, the level of the effort far exceeds that which it was possible to attain even as recently as seven years ago. In a sense, IGY really set the stage for IQSY.

Scientific Objectives. The broad objective of IGY was to study every aspect of the earth as a planet, including its environment. Consequently, in addition to investigating the properties of the earth's interior (seismology, latitude and longitude, and gravimetry) and studying its surface characteristics (oceanography and glaciology), the program embraced observations in and beyond the earth's atmosphere, including investigations of sun-earth relationships (meteorology, nuclear radiations, geomagnetism, ionosphere, aurora and airglow, cosmic rays, and solar activity). And, to increase basic knowledge about the solar influences acting upon the earth, IGY was planned to cover a period at or near solar maximum (see also GEOPHYSICS and SOLAR PHYSICS).

IGY was *not* the first international effort, however. There had been similar undertakings before: The First International Polar Year, 1882–1883, and the Second International Polar Year, 1932–1933. The recent enterprises are distinctive for other reasons; they involve the first extensive *in situ* probing of the earth's envelope, thanks to the subsequent development of rockets and, most recently, of satellites.

One motivation for IQSY was to utilize to the fullest extent the remarkable new technology and facilities, spawned by IGY, that have now matured to a previously undreamed of stage of development. Not only are vehicles now available for conducting highly sophisticated experiments in space, but other new techniques have added new dimensions greatly extending the possibilities which existed even as recently as seven years ago. Furthermore, logistic capabilities have materialized that make it feasible to carry out large-scale programs in crucial but previously inaccessible regions of the globe.

Antarctica, as a locale for research, deserves special mention. Not only is it important for studying upper atmosphere phenomena on a global scale, but it is vital in the conduct of all studies of upper atmosphere phenomena. The key to the preeminent role of the polar regions in the upper atmospheric research is the terrestrial magnetic field. Electrically-charged particles approaching the earth can, unless they are endowed with higher energies, arrive only near the geomagnetic poles. Consequently, the lower energy protons and heavier atomic nuclei that are sometimes produced by the sun can be observed only in these regions. Similarly, auroral effects, as well as other phenomena relating to magnetic lines of force that extend to very great distances from the earth's surface, can be observed only near the geomagnetic poles.

The purpose of IQSY is threefold. Some of the studies that are being conducted are feasible, or are best undertaken, only at the time of solar minimum. Others are concerned with observing in detail isolated solar events, uncomplicated by

the superposition in time of a number of concurrent outbursts. Finally, some investigations are providing data characteristic of solar minimum conditions for comparison with those obtained previously during solar maximum.

Actually, at sunspot minimum, solar outbursts do not cease completely, but when they do occur, their effects can be observed under relatively "clean" conditions, free from confusing interferences. Hence, the entire sequence of events associated with a single outburst can be followed. Furthermore, smaller effects, previously lost in the high-level background of activity, are discernible. Thus, many IQSY projects are taking advantage of this period to make observations with greatly increased "amplification".

The IQSY Program. The IQSY program is organized in eight disciplines, although in most cases there are overlapping interests, and no sharp boundary line is drawn. The disciplines involved are the following:

Meteorology. The upper 10 per cent of the atmosphere is emphasized. This is the region in which solar influences are propagated to the lower atmosphere. Every available technique, including the use of rockets and satellites, is being utilized for determining the meteorological parameters that are relevant to atmospheric energy transfer (see METEOROLOGY).

Geomagnetism. A detailed study of the earth's main magnetic field is being conducted. Magnetic disturbances, which are related to many geophysical phenomena, are observed by an extensive network of ground-based observatories. Rockets and satellites are also providing important information extending through the magnetosphere, the magnetopause (termination of the earth's magnetic field), and the transition region between the interplanetary and terrestrial fields (see GEOPHYSICS).

Aurora. Instrumentation carried aloft by balloons, rockets, and satellites is being employed to determine the characteristics of aurora-producing particles and the mechanisms of particle acceleration and precipitation, supplementing studies based upon observations from the ground (see AURORA AND AIRGLOW).

Airglow. Airglow emissions yield important information about the structure and chemical composition of the upper atmosphere. The geocorona, a ring of hydrogen that girds the earth and is detectable through the Lyman- α light that it emits, is of special importance because of its response to the flow past the earth of the solar wind (see AURORA AND AIRGLOW).

Ionospheric Physics and Radioastronomy. The ionosphere, well known because of its association with radio transmission, is being probed by a variety of means both from above and below. Observations at conjugate points (opposite ends of a magnetic line of force) on the earth's surface are also being made. Advantage is taken of the decreased opacity of the ionosphere for extending radioastronomical observations of galactic and extragalactic sources to lower frequencies, and radio emissions from the planets are also being

studied (see IONOSPHERE and RADIOASTRONOMY).

The Sun and the Interplanetary Medium. Constant surveillance of the sun is being maintained to detect disturbances over a broad spectrum of wavelengths. Solar structure is being investigated with new techniques, and magnetic fields originating at the sun and carried into space by the corpuscular radiation which constitutes the solar wind are being measured directly. The interplanetary medium is being probed both by direct and indirect methods (see SOLAR PHYSICS and SPACE PHYSICS).

Cosmic Rays and Geomagnetically Trapped Radiation. The lower energy cosmic rays, shielded by clouds of magnetized plasma emitted by the sun, can be observed only at solar minimum. Various characteristics of both the galactic cosmic rays, and of solar-produced particles, are being investigated by a variety of means (see COSMIC RAYS and RADIATION BELTS).

Aeronomy. *In situ* measurements with rocket and satellite vehicles are adding to our knowledge of the physics and chemistry of the upper atmosphere. By obtaining a complete description during quiet conditions, a base line will be provided for comparison with events occurring in association with solar disturbances (see PLANETARY ATMOSPHERES).

In the United States, a Committee of the National Academy of Sciences, the U.S. Committee for IQSY, is responsible for development of the program. The members are: R. G. Athay (solar activity), J. W. Chamberlain (aurora and airglow), H. Friedman (aeronomy), J. Kaplan (member-at-large), W. W. Kellogg (meteorology), P. Meyer (cosmic rays), H. Odishaw (ex officio), M. A. Pomerantz (chairman), E. Dyer (executive secretary), M. A. Tuve (ex officio), E. H. Vestine (geomagnetism), and A. H. Waynick (ionospheric physics). Within the government, the National Science Foundation was designated by the president as the responsible agency for coordinating and implementing the program, and for correlating regular activities of the government that relate to the program. Robert Fleischer, of the Office of Atmospheric Sciences, is NSF coordinator for IQSY.

The Special Committee for IQSY, established by the International Council of Scientific Unions (ICSU), with a Secretariat in London, formulated and is executing the detailed plans of this great cooperative venture in which scientists of seventy-one nations are participating. The Bureau is comprised of the following members: W. J. G. Beynon, president; M. A. Pomerantz, N. V. Pushkov, and G. Righini, vice presidents; and C. M. Minnis, secretary.

MARTIN A. POMERANTZ

References

1. A discussion of the history and development of the IGY is given in "Annals of the IGY," New York, Pergamon Press, 1959, Vol. I; accounts of the five general assemblies of the IGY are to be found in volumes 2a, 2b, and 10. There have been many

- excellent general books and articles on the IGY: see, for example, Chapman, S., "Year of Discovery," Ann Arbor, University of Michigan Press, 1959 and Sullivan, W., "Assault on the Unknown," New York, McGraw-Hill Book Co., 1961.
2. Pomerantz, Martin A., "International Years of the Quiet Sun 1964-65," *Science*, **142**, 3596 (1963).
 3. The "IQSY Calendar" is published by and may be obtained from International Scientific Radio Union, 7 Place Danco, Brussels 18. Copies are also available from the IQSY Secretariat. The calendar is reproduced in *IG Bull.*, No. 74 (1963).
 4. A comprehensive bibliography of U.S. contributions to the IGY is available: "United States IGY Bibliography, 1953-60," *Nat. Acad. Sci. Publ.*, **1087** (1963). An international bibliography has been published in "Annals of the IGY," New York, Pergamon Press.

IONIZATION

Ionization is the name given to any process by which a net electrical charge may be imparted to an atom or group of atoms. In the case of liquid solvents, molecules or ionic salts become dissociated to form positive and negative ions. This ionization process is known as *electrolysis*, and the name *electrolyte* is given to the solute or to the conducting solution. The study of electrolysis is embodied in the subject of *electrochemistry*. Of great interest in recent years has been the study of ionized gases. Rockets, hypersonic flight, and space physics have spurred investigations of plasmas, shock waves and high-temperature chemical processes arising in a variety of terrestrial and celestial phenomena. Indeed, ionized gases make up a major portion of all matter in the universe. Our chronic need for new energy sources has transformed the speculation of a controlled thermonuclear fusion reaction into one of the greatest research efforts in history. These and other considerations have induced a vigorous growth in the study of ionization phenomena.

Electrolytes. The degree of ionization found in electrolytes is highly variable and depends upon the solute, the solvent, and the interaction between them. *Weak* electrolytes, such as many organic compounds, are solutes which are barely dissociated into ions except in the limit of infinite dilution. *Strong* electrolytes are highly dissociated at any concentration. Ions formed in solution may bear one or several electronic charges. The *electrochemical equivalent weight* is the atomic weight divided by the number of charges carried by the ion. If electrodes are placed in an electrolyte and a current flows in the external circuit, the ions with positive charges, called *cations* (cathode + ions), will migrate toward the negative electrode (cathode). Those ions possessing a negative charge (*anions*) will migrate to the positive electrode (anode). The ions arriving at the cathode are neutralized by the acquisition of electrons; the atoms or molecules thus formed may then be evolved as a gas or retained as a deposit on the electrode. The cations are said

to undergo *reduction*. Likewise, the anions experience a loss of electrons at the anode; this process is called *oxidation*. The quantity of electricity required to deposit one gram equivalent is called the faraday, in honor of Michael Faraday (1791-1867). Based on the physical scale of atomic weights, the faraday is numerically equal to 96,520 coulombs per equivalent.

As for any solid conductor of uniform cross sectional area A and length l , the electrical resistance R of an electrolyte is given by $l/\kappa A$. The conductivity κ is independent of geometrical shape and size, and bears the units $(\text{ohm cm})^{-1}$. Of greater importance in the study of electrolytes is the *equivalent conductivity*:

$$\Lambda = \kappa/C$$

where C is the concentration of the solute in equivalents per cubic centimeter. Plots of Λ as a function of concentration show very different behavior for weak and strong electrolytes. The latter exhibit a limiting value of Λ as C diminishes to zero, while the weak electrolytes do not. Such behavior provides insight into the nature of the ions, their mobilities and their interactions with the solvent material.

Formation of Gaseous Ions. Studies in 1895 by J. J. Thomson of the effects of newly discovered x-rays on gases marked the beginning of a series of experiments which established the existence of the electron and clarified many questions on the nature of atomic structure.

Just as in electrolytes, both positive and negative ions may exist in an ionized gas. In addition to the ions, the presence of free electrons may profoundly influence the character of the gas. Negative ions may be formed by the attachment of free electrons to a neutral atom or molecule, by the dissociation of a neutral molecule into positive and negative fragments, or by electron transfer upon collision of two neutral atoms or molecules. Positive ions may be formed by dissociation, charge transfer, neutral-particle or electron collisions, or by photoabsorption (the absorption of electromagnetic radiation). Still another mechanism for the formation of ions is the emission of a nuclear particle, such as beta decay. Several of these processes are discussed below.

Photoionization. Photoabsorption leading to excitation and ionization is of interest because of its significance in astrophysics and geophysics. The ionosphere is constituted of molecular and atomic ions which result from the absorption of solar ultraviolet and x-radiation. The frequency ν of the electromagnetic radiation giving rise to ionization must satisfy the relation $h\nu \geq V$, where h is Planck's constant and V is the *ionization potential*. The latter is defined as the energy required to remove completely an electron from an atom or molecule in the ground state, leaving the resulting ion in its lowest state. Photons having energy less than V may be absorbed by atoms or molecules, giving rise to excitation of internal states or perhaps molecular dissociation, or both.

Many laboratory investigations of photoabsorption have been performed. One type of experiment requires the measurement of the absorption coefficient, α , of a photon beam:

$$I = I_0 e^{-\alpha x}$$

in which I_0 is the initial intensity of the beam and I is the intensity after the beam has traversed a distance x in the absorbing gas. The absorption coefficient may then be studied as a function of photon energy (i.e., wavelength). It can be expressed in terms of a microscopic cross section for absorption or for ionization as a function of incident wavelength. This cross section curve for photoionization usually exhibits a sharp peak at the ionization threshold. Only a single, outermost electron is ejected from an atom which absorbs an ultraviolet photon. An x-ray photon generally will eject a more tightly bound electron from one of the atom's inner shells.

Other types of experiments utilize photoionization to study the deionization process for the ions thus formed. Of major importance are such processes as electron-ion recombination:



and ion-ion recombination if both charged species are ions.

Ionization by Heavy-particle Collisions. By heavy particles is meant both atoms and molecules and their ions, ranging in mass from the hydrogen atomic ion (proton) to very heavy molecular systems of large atomic number. When two heavy particles collide with sufficient energy, one or more electrons may be ejected from either or both particles. In experimental work, the *target* molecules or atoms are in the form of a low-density gas having an energy corresponding to room temperature and usually negligible compared to the energy of the projectile particles. The latter are usually obtained through ionization of a selected gas in an ion source, and acceleration through a large electric potential difference E . Regardless of the mass of the projectile particles, their kinetic energy will be equal numerically to E electron volts, if the potential difference E is in volts and if the particles carry but one elementary charge. One electron volt (eV) is equal to 1.6×10^{-12} erg.

If a beam thus formed with an intensity of B particles per second is incident on a target chamber of area A containing N target particles as a low-density gas, the electron ionization current i which is released is given by

$$i = BN\sigma/A \text{ electrons/sec.}$$

This equation defines the effective ionization cross section σ . For energies at which multiple ionization is improbable, σ approaches the cross section for singly charged ions. Often the distinction is made between ionization of the target particle and the beam particle. Ionization of the latter is referred to as stripping.

In the quantitative description of heavy-particle collisions, one usually introduces the concept of

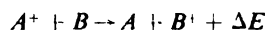
the (a) laboratory and (b) center-of-mass (CM) coordinate systems. The laboratory system is used to describe the motion of the particles as would be viewed by an observer standing at rest in the laboratory. The origin of the CM system moves with the center-of-mass of the two-particle system. If m and M are the masses of the projectile and target particles, the former moving with a velocity v much greater than the target velocity, the center-of-mass velocity V_c is

$$V_c = mv/(m + M) = \mu v/M$$

The latter relation defines μ , the reduced mass. The kinetic energy in the CM system is $\frac{1}{2}\mu v^2$, which is the projectile energy multiplied by $M/(M + m)$. For the case in which target and projectile are identical, the kinetic energy in the CM system is half that of the projectile. Using the law of conservation of linear momentum, one may show that the kinetic energy in the CM system is the maximum energy available for excitation and/or ionization.

Collisions in general are classified as (a) elastic in which no changes in internal states occur, and for which kinetic energy is conserved, and (b) inelastic, in which a part of the kinetic energy is converted to internal energy. A *superelastic* collision is one in which internal energy is transformed into kinetic energy. Because of the extreme complexity encountered in quantum-theoretical calculations of ionization cross sections for heavy-particle collisions, very little progress has been made in this important area of collision theory.

Ion-neutral collisions may give rise to free electrons, or simply *charge exchange* which, in its simplest form, is expressed by the equation.



The neutral particle B has been ionized, but the electron has transferred to the incident ion, neutralizing it. The energy ΔE released in this process is the difference between the ionization potentials of the neutral particles A and B . For the case in which A and B are identical, $\Delta E = 0$ and the process is called *symmetric resonant* charge transfer. At low ion beam energies, this transfer proceeds with a large cross section.

Ionization by Electron Impact. Of great importance in atomic physics is the ionization produced by collimated beams of electrons incident on heavy particles. Whether the target particles are atoms or molecules, the ionization is usually by the removal of single electrons from the outermost shell, as in photoionization. As a function of the incident electron energy, the ionization cross section rises rapidly from zero for energies just below the ionization potential and increases to a maximum value in the neighborhood of 50 to 100 eV; thereafter it decreases slowly and monotonically with increasing electron energy. Since an electron with a given energy travels at much higher speed than does a heavy projectile of the same energy, the electron collision induces a much more rapid perturbation on the target's

orbital electrons. Thus, a larger ionization cross section at low energies is to be anticipated.

Much of the definitive work on electron impact ionization was performed in the 1930's by Tate, Smith, and Bleakney (cf. reference 4).

Collective Processes. If, as in a glow discharge, a large number of charged particles are created, the collective interactions of these particles with each other and with external fields may permit the charged fluid to exhibit very unusual and distinct properties. An ionized gas possessing both positive and negative charges is called a *plasma* if the distance over which the gas can have an appreciable departure from charge neutrality is small compared to the dimensions of the gas. This distance is described by the *Debye-Hückel radius*, a quantity borrowed from the theory of strong electrolytes, which characterizes the decay of the shielded Coulomb potential surrounding the ionized particles of the fluid. If the charge neutrality in a plasma is disturbed in some manner, the electrons will be forced to oscillate about their equilibrium positions in simple harmonic motion with a frequency characterized by the electron density. Longitudinal oscillation of the ions and electrons as a whole constitutes another type of motion called *ion-acoustical waves*. *Hydro-magnetic waves*, which appear in the presence of a magnetic field, are still another form of motion not observed in an unionized medium. The description of such phenomena goes beyond the scope of the ionization process.

ROBERT C. AMME

References

1. Condon, E. U., and Odishaw, H., "Handbook of Physics," New York, McGraw-Hill Book Co., 1958.
2. Spitzer, L., "Physics of Fully Ionized Gases," Interscience Tracts on Physics and Astronomy, New York, Interscience Publishers, 1956.
3. Loeb, L. B., "Basic Processes of Gaseous Electronics," Berkeley, University of California Press, 1955.
4. McDaniel, E. W., "Collision Phenomena in Ionized Gases," New York, John Wiley & Sons, 1964.
5. Hasted, J. B., "Physics of Atomic Collisions," Washington D.C., Butterworth, 1964.

Cross-references: COLLISIONS OF PARTICLES, CROSS SECTIONS AND STOPPING POWER, ELECTRICAL DISCHARGES IN GASES, ELECTROCHEMISTRY, IONOSPHERE, MAGNETO-FLUID-MECHANICS, PLASMAS.

IONOSPHERE

The ionosphere is the gas of charged, thermal particles which forms a fraction of the earth's atmosphere between about 50 km above the surface and the outer limit of the atmosphere. The term "ionosphere" is also used to mean the atmospheric region in which the ionosphere lies. Other planets have ionospheres too, although,

of course, our knowledge of them is still almost entirely theoretical. The ionosphere is also sometimes defined as "the part of the earth's upper atmosphere where ions and electrons are present in quantities sufficient to affect the propagation of radio waves." Indeed, the ionosphere has been studied primarily by its effects on radio waves, although *in situ* measurements from rockets and satellites are playing an increasingly important role. This definition also emphasizes the practical importance of the ionosphere: it refracts and reflects radio waves and so makes possible beyond-the-horizon radio propagation.

The ionosphere is divided into six regions or layers which are called, from bottom to top, the C, D, E, and F regions, the heliosphere, and the protonosphere. The F region is subdivided into the F1 and F2 layers, and the C region was, until recently, considered to be the lower part of the D region. These regions or layers are more or less identifiable on a graph of electron concentration vs. height, but they are properly defined in terms of the processes by which they are formed, and not by the height ranges in which they lie. Nevertheless, their typical heights are useful. The lower limit of the C region is conventionally 50 km. The boundary between the C and D regions is at about 70 km; between the D and E regions, at about 90 km; between the E and F regions, at 120 to 140 km, and between the F1 and F2 layers, at about 200 km. The heights of the boundaries between the F region and the heliosphere and between the heliosphere and the protonosphere are poorly observed, but they are known to increase with increasing atmospheric temperature. Thus, the lowest heights, which may be as low as 500 km, occur at night and during periods of minimum solar and geomagnetic activity. The highest heights, which occur during the day, and during periods of maximum solar and geomagnetic activity, may be very much greater. The outer limit of the protonosphere (the magnetopause), which is also the outer limit of the ionosphere and the earth's atmosphere, is always at least several earth radii from the earth.

In the C and D regions, the positive ions are probably mostly NO^+ , and the negative charge is carried partly by electrons, partly by O_2^- , and partly by some other species of negative ions. In all of the higher regions, the negative charge is carried almost entirely by electrons. In the E and F1 regions, the positive ions are mostly O_2^+ and NO^+ ; in the F2 layer, O^+ ; in the heliosphere, He^+ ; and in the protonosphere, H^+ (protons). Throughout, the ionosphere is very nearly electrically neutral; i.e., the total positive and negative charge concentrations are very nearly equal.

Typical values of the electron concentration at noon are: C region, 10^2 electrons/cm³; D region, 10^3 to 10^4 ; E and F1 regions, 10^5 ; F2 layer, 10^6 ; heliosphere, 10^4 ; protonosphere, 10^4 to 10^5 . At night the concentrations in all of the regions tend to be smaller; in the D and E regions, by a factor of 10. In all of the regions, the electron

concentration has complex dependences on the time of day, season, latitude, and solar and geomagnetic activity. The variations of the F2 layer are particularly large and well-observed: its concentration ranges from 10^4 to 10^7 electrons/cm³ under different conditions.

Superimposed upon the rather smooth ionization concentration of the ionospheric regions are many kinds of spatial irregularities, which are loosely grouped into "Sporadic E" in the E region and "Spread F" in the F region. Sporadic E is usually found in thin horizontal sheets. One of the commonest types consists of a slab only a few hundred meters thick and tens of kilometers in horizontal extent in which the electron density may be as much as a factor of ten larger than in the ambient E region. In the F region, on the other hand, the irregularities are filaments aligned along the lines of force of the earth's magnetic field. They are from a few meters to a few kilometers across and they extend along the field lines from the lower F2 region indefinitely upwards, even to the opposite hemisphere. The individual filaments appear to be arranged in east-west sheets. The electron concentration in the F-region irregularities is usually no more than a few per cent greater or less than that in the ambient ionosphere.

The observational and theoretical understanding of these irregularities has proceeded rapidly in recent years, but semiquantitative theories exist only for two or three kinds of Sporadic E.

The charged particles which constitute the ionosphere are formed by ionization of the atoms and molecules of the ambient un-ionized atmosphere. Under geomagnetically quiet conditions, the C region is caused by the ionization of N₂ and O₂ by galactic cosmic rays, but all of the other regions are caused by solar photons in the extreme ultraviolet and x-ray part of the spectrum.

The quiet D region is formed by the ionization of NO—a trace in the atmosphere—by the solar Lyman- α line at 1261Å. The quiet E and F regions are formed by the ionization of all of the atmospheric constituents there, i.e., N₂, O₂, and O (which is formed from O₂ by photodissociation), by solar photons in different bands between about 10 and 1027Å. The heliosphere is formed by the ionization of He by wavelengths between 165 and 504Å. Photoionization of H by wavelengths between 165 and 911Å makes some contribution to the protonosphere, but another process, to be discussed later, is dominant.

The rate of ionization is occasionally considerably increased by several kinds of disturbances. The C and lower D regions in the polar regions are enhanced by ionization by solar protons during Polar Cap Absorption events. Enhancements of the D, E, and F regions by ionization by fast electrons (called Auroral Absorption events) are frequent in the polar regions but occur at lower latitudes only during geomagnetic disturbances. The D, E, and F regions are also

enhanced by solar photons emitted during solar flares; these enhancements are called Sudden Ionospheric Disturbances. Enhancements of ionization in the D region cause increased absorption of radio waves, with deleterious effects on radio communication.

The positive and negative charges eventually recombine to form neutral particles. However, some of the species of ions—especially the atomic ions—are more likely to undergo reactions which form other species of ions than they are to recombine with electrons. Thus, the protons in the protonosphere are lost by transferring their charge to O atoms to form O⁺ ions. Since the ionization potentials of H and O are almost equal, this reaction can also proceed in the other direction to form H⁺ from O⁺. This is the principal source of ionization in the protonosphere. The He⁺ in the heliosphere is lost by chemical reactions with N₂ and O₂ which form N⁺ and O⁺. These atomic ions are lost by further chemical reactions with N₂ and O₂ which form NO⁺ and O₂⁺. These molecular ions can then recombine with electrons by a process (dissociative recombination) which is about 10⁵ times faster than the recombination of atomic ions (by radiative recombination). Also, in the C and D regions the electrons can attach themselves to neutral particles to form negative ions, which can then either recombine with positive ions or be detached.

Above about 200 km, that is, in the F2 layer, the heliosphere and the protonosphere, plasma diffusion plays an important role in determining the distribution of ionization. The forces which cause this diffusion are primarily gravity, concentration gradients, and electric fields. Above the peak of the F2 layer, at about 300 km, the ionospheric plasma tends to be in diffusive equilibrium, and the electron concentration decreases exponentially with height, with a logarithmic decrement which is proportional to the mass of the ions.

The complex variations of the electron concentration in the ionospheric regions are the results of the interplay among the processes of ionization, recombination, and diffusion, each of which itself has a complex dependence on time of day, season, latitude, and solar and geomagnetic activity. Although the general nature of these various dependences is known, their rates are not yet known well enough to form a satisfactory quantitative model of the ionosphere.

T. E. VANZANDT

References

- Ratcliffe, J. A., and Weekes, K., Ch. 9, pp. 377–470, in Ratcliffe, J. A., Ed., "Physics of the Upper Atmosphere," New York, Academic Press, 1960.
- VanZandt, T. E., Cohen, R., and Reid, G. C., in Campbell, W. H., and Matsushita, S., Eds., "Physics of Geomagnetic Phenomena," New York, Academic Press, 1965.

Cross-references: AURORA AND AIRGLOW, IONIZATION, PLANETARY ATMOSPHERES, SPACE PHYSICS.

IRRADIATION, DISPLACED ATOMS

High-energy particles interact with the atoms of a solid in several ways and thereby produce disturbances in the atomic and electronic structure of the solid. The practical importance of such interactions is that many physical properties are very sensitive to the disturbances produced by radiation. Drastic changes, often deleterious and therefore referred to as radiation damage, may occur in such properties of practical importance as dimensional stability, mechanical and electrical properties, thermal conductivity, etc. The scientific importance of the field arises from the fact that the study of radiation effects leads to new and valuable insight into the properties of imperfections in solids. Irradiation with energetic particles has become a powerful tool of solid-state research, since a large number of imperfections can be introduced into a solid in a reasonably well-controlled manner.

The most important basic processes arising from the interaction of high-energy particles with solids may be classified as follows: (1) production of displaced and excited electrons, i.e., ionization; (2) production of displaced atoms by direct collision; and (3) production of fission and thermal spikes. In some cases transmutation effects also have to be taken into account. Attention is focused here on the production and nature of displaced atoms. Neglect of ionization is realistic for metals whose electrical conductivity is high, because in such metals any ionization is so rapidly neutralized that ionization effects are not observable. In insulators, ionization effects are of primary importance, and in semiconductors, both ionization effects and displacement production are important. Ionization effects are discussed under such entries as COLOR CENTERS; RADIATION CHEMISTRY; RADIATION, IONIZING, BASIC INTERACTIONS. Fission and thermal spikes, rather complex disturbances, are important in materials of high atomic number irradiated with massive particles but will not be further discussed here (see references 3 and 4).

If a bombarding particle makes an elastic collision with an atom and transfers to it an amount of energy larger than the displacement threshold energy (typically about 25 eV), the atom will be displaced from its lattice position. In most cases the displaced atom, or knock-on, has enough recoil energy to travel a few atomic distances from its initial position, either directly through the lattice or via a series of replacement collisions (see below), before coming to rest in an interstitial position. Thus, the fundamental displacement pair is produced: the displaced atom or interstitial, and the lattice site which was left empty, the vacancy. A complete theory of defect production at all energies is not yet at hand, but a great deal of insight can be obtained from theoretical dynamic studies at rather low energies made with high-speed computers. In Fig. 1 three important processes that occur during displacement production are illustrated. Atom A, the primary knock-on, is assumed to have been struck

by a bombarding particle and to have been given an initial energy of 40 eV in the direction indicated by the arrow at A. The lines in this Figure are the paths followed by the individual particles. By subsequent collisions, an interstitial has been produced at location D and a vacancy left behind at position A. Replacement collisions have occurred at locations B and C, where atom A replaced atom B and, in turn, atom B replaced atom C. Focusing collisions, preferential propagation of energy along rows of close-packed atoms, are also clearly seen in the figure. At higher energies the focusing collisions transport matter as well as energy and thus create interstitial atoms, after a series of replacement collisions, at some distance from the original point of impact. The picture becomes considerably more complex at still higher energies. Thus, results of experiments involving irradiations with electrons or gamma rays in the few-MeV range are much easier to interpret than the more complex damage resulting from irradiation with neutrons (in a reactor) or with heavy charged particles.

The displacement process is clearly quite complex, and no complete quantitative theory has been formulated. The average number of displaced atoms can be estimated by means of a simple model based on binary collisions. Comparison with experiment shows rather good agreement in the case of metals irradiated with electrons. Upon heavy particle or reactor irradiation, the experimentally observed concentration of displaced atoms is less than that predicted by theory by a factor of 5 to 10. A rather crude, but convenient, number to remember is that in most metals one atomic per cent vacancy-interstitial pairs are produced by 10^{20} neutrons/cm² in a reactor exposure to fast (epithermal) neutrons.

A convenient and accurate measure of defect concentration is the low-temperature electrical resistivity, or residual resistivity, of a metal. Much of the fundamental information available has been obtained by this technique. Other physical property changes, however, are far more important from a practical standpoint. For example, metals generally harden considerably upon reactor irradiation. The increase in critical shear stress is usually accompanied by a reduction in ductility leading to increased brittleness. Such changes are intricate since they involve the interaction of the radiation-induced defects with static and moving dislocations, and a full interpretation of the experiments is not at hand. As another example, graphite exhibits a large increase in volume upon reactor irradiation. It is quite clear in this case that the radiation-induced interstitials lodge between the graphite planes and push these planes apart.

A very important fundamental and practical property of the radiation-induced defects is their mobility. The defects can migrate in the solid from one position to another by surmounting an energy barrier, i.e., their migration is characterized by an activation energy. Since an irradiated crystal is not in thermodynamic equilibrium, it will tend to revert to its stable unirradiated form

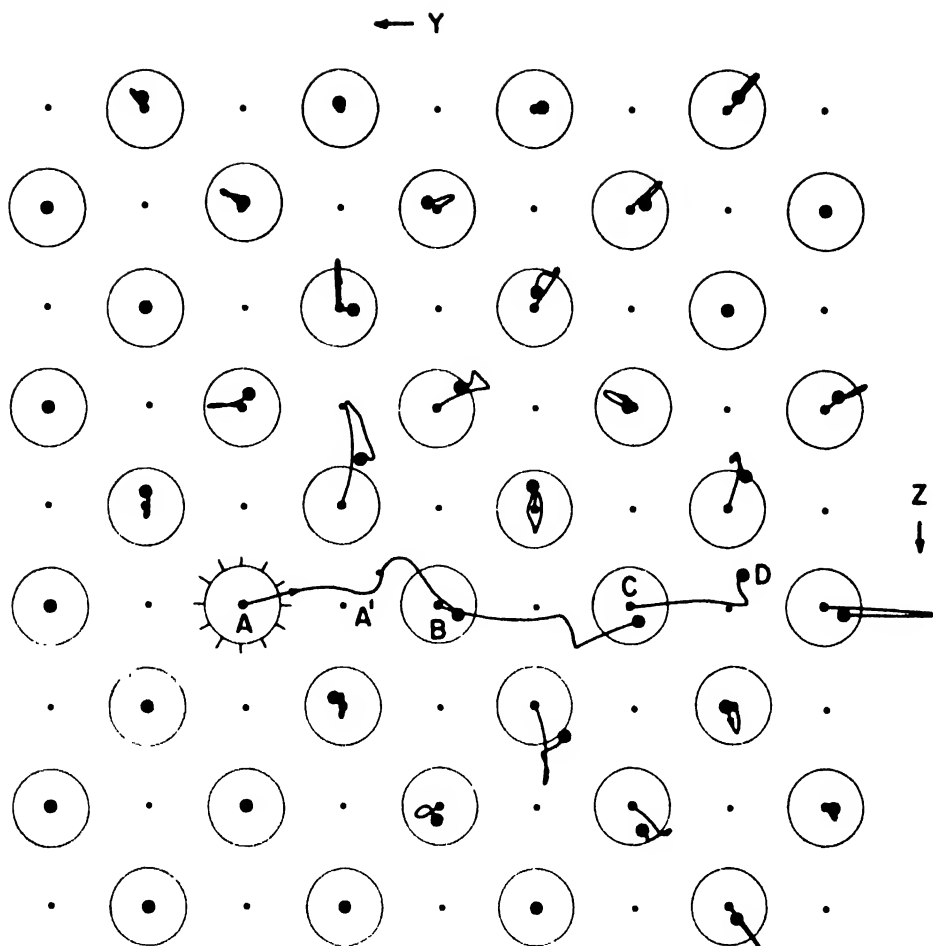


Fig. 1. Atomic paths, calculated dynamically, produced by a 40-eV knock-on in copper. The primary knock-on collision was at lattice position A (from reference 6).

at temperatures high enough for the defects to be mobile, i.e., the crystal can be annealed. The kinetics of the annealing process can be quite complex, since usually several competing processes are going on such as vacancy-interstitial annihilation, defect migration to external and internal surfaces, defect aggregation, etc. From the fundamental standpoint, a great deal can be learned about the characteristics of the defects from detailed annealing experiments. The practical importance of annealing is that by raising the temperature undesirable radiation damage can be minimized. The situation is more complicated in alloys, since the migration of defects is equivalent to diffusion, which may result in metallurgical changes. Indeed, radiation-enhanced diffusion, nucleation, precipitation, and phase transformation have been observed.

Ionic crystals are discussed elsewhere. In semiconductors, the fundamental displacement process is very similar to the one already described. However, the radiation-induced defects in semiconductors have localized energy states associated

with them which alter the concentration and mobility of charge carriers and thereby lead to drastic and important changes in the electronic properties. Some of the defect energy states are occupied by electrons and are therefore donors, while others are vacant and act as acceptors. In some semiconducting crystals, the radiation-induced changes go very far indeed. For example, germanium and gallium antimonide are converted from *n*-type to *p*-type upon irradiation. The sensitivity of semiconductor devices to radiation is of great practical importance in many space applications (solar cells, electronic control and detection devices, etc.).

G. J. DIENES

References

1. Damask, A. C., and Dienes, G. J., "Point Defects in Metals," New York, Gordon and Breach, 1963.
2. "Radiation Damage in Solids," Vols. I-III, Vienna, International Atomic Energy Agency, 1962.

3. Billington, D. S., and Crawford, J. H., "Radiation Damage in Solids," Princeton, N. J., Princeton University Press, 1961.
4. Dienes, G. J., and Vineyard, G. H., "Radiation Effects in Solids," New York, Interscience Publishers, 1957.
5. Seitz, F., and Koehler, J. S., "Displacement of Atoms During Irradiation", in Seitz, F., and Turnbull, D., Eds., "Solid State Physics", Vol. II, pp. 307-449, New York, Academic Press, 1956.
6. Gibson, J. B., Goland, A. N., Milgram, M., and Vineyard, G. H., *Phys. Rev.*, **120**, 1229 (1960).

Cross-references: COLOR CENTERS; NUCLEAR REACTIONS; RADIATION CHEMISTRY; RADIATION, IONIZING, BASIC INTERACTIONS; SEMICONDUCTORS; SOLID-STATE PHYSICS; SOLID-STATE THEORY.

ISOTOPES

The word "isotopes," stemming from the Greek words, *isos* (same) and *topos* (place), refers to atoms of an element which have differing masses. The term was first proposed by Soddy in 1913 to designate substances having different atomic weights and yet having chemical properties which were so closely allied that no chemical method was effective in producing a separation. Hence, Soddy suggested that they were chemically identical; i.e., they occupied the same place in the periodic table. In 1905, Boldwood noted the presence of lead in uranium minerals and suggested that this might be the end product of the uranium series. As a result of the study of the relation of lead to uranium in a large number of minerals, this view was generally adopted. Soddy concluded that the end products of the uranium and thorium series should be lead with isotopic weights of 206 and 208, respectively, whereas ordinary lead has an atomic weight of 207.2. Soddy and Hyman reported in 1914 that the atomic weight of lead as found in thorite (consisting mainly of thorium, 1 to 2 per cent uranium, and 0.4 per cent lead) indeed had a slightly greater atomic weight than that of ordinary lead.

Some elements such as sodium and cobalt are mononuclidic; i.e., all of the natural stable atoms have the same mass. Others have many natural stable isotopes, as for example, tin, which is made up of atoms of ten different masses as it occurs in nature. Since the atoms in multinuclidic elements are chemically identical and have the same number of protons in the nucleus, the varying masses are accounted for by the variable number of neutrons in the nucleus.

Isotopes are divided into two groups: stable and radioactive (unstable). The total known isotopes number about 1500, of which 280 are stable and the balance are radioactive, having a transient existence ranging from millionths of a second to millions of years. Radioisotopes undergo transformation, or decay, emitting alpha, beta, gamma, or x-radiations during their return to a stable condition (see RADIOACTIVITY).

Isotopes, both stable and radioactive, have grown in importance to science and technology in

the last 40 years. Since atoms can be marked by their radioactivity, or in some cases by an atypical isotopic composition, the elements can be traced, a procedure of great value in physical and biological science, technology, and medical diagnosis. Further, radioisotopes emit corpuscular and/or electromagnetic radiations that can be used to probe into and through matter, affect it chemically and physically, produce heat and light, kill microorganisms, and perform many other tasks of benefit in today's complex industrial society.

Stable Isotopes. By 1900, physicists had found that positively charged particles formed by the passage of an electric discharge through an evacuated tube consisted of molecular ions of the gas present in the tube. Deflection of these positive rays by electric and magnetic fields offered a sensitive tool to study gaseous elements. By allowing the rays from a given element to fall on a photographic plate, a series of parabolic streaks were observed, each corresponding to a definite value of mass-to-charge ratio (m/e). Positive ray photographs of neon (atomic weight 20.2) obtained by Thomson exhibited, among other things, a heavy neon line and a faint line at 22. In an effort to elucidate the situation, Thomson's assistant, Aston, passed neon gas through a porous pipe-clay tube repeatedly and was able to show a significant alteration in the atomic weight of the two extreme fractions. This alteration was reflected in changes in the relative brightness of the two lines in subsequent positive ray analyses.

Aston proceeded to redesign the positive ray apparatus so that the particles having the same mass were brought to a focus to produce a sharp line rather than a parabola; the resulting instrument was called a mass spectrograph. With this instrument, Aston was able to confirm the finding that neon exists in at least two forms (atomic weights of 20 and 22) and that the proportions appeared to be 10:1, giving an average atomic weight of 20.2 to neon. Aston next analyzed chlorine and also found that this gave two lines, corresponding to 35 and 37.

The pioneering work of Aston and Dempster in 1918-19 with the electromagnetic mass spectrometer is the historical starting point for separation and study of the isotopes of the elements. The electromagnetic separation of isotopes is relatively simple in principle (see MASS SPECTROMETRY).

Production of Stable Isotopes. There are a number of possible ways of separating isotopes using electromagnetic principles. However, the large-scale mass spectrometer known as a *calutron* is the device now used almost exclusively. Within a tank maintained at high vacuum, ions of an element are produced by vaporization at high temperature, sometimes assisted by a chemical agent such as carbon tetrachloride (chlorination). The ions are accelerated by an electric potential and projected as a beam across a magnetic field. While the electromagnetic forces acting to curve the path of the ions are essentially the same on each of the ions, they curve into different paths because the masses of the several isotopes are

TABLE I. TABLE OF RELATIVE ATOMIC WEIGHTS 1961

Based on the Atomic Mass of C^{12} --- 12

Order of Atomic Number			
Atomic Number	Name	Symbol	Atomic Weight
	Hydrogen	H	1.00797
2	Helium	He	4.0026
3	Lithium	Li	6.939
4	Beryllium	Be	9.0122
5	Boron	B	10.811
6	Carbon	C	12.01115
7	Nitrogen	N	14.0067
8	Oxygen	O	15.9994
9	Fluorine	F	18.9984
10	Neon	Ne	20.183
11	Sodium	Na	22.9898
12	Magnesium	Mg	24.312
13	Aluminum	Al	26.9815
14	Silicon	Si	28.086
15	Phosphorus	P	30.9738
16	Sulfur	S	32.064
17	Chlorine	Cl	35.453
18	Argon	Ar	39.948
19	Potassium	K	39.102
20	Calcium	Ca	40.08
21	Scandium	Sc	44.956
22	Titanium	Ti	47.90
23	Vanadium	V	50.942
24	Chromium	Cr	51.996
25	Manganese	Mn	54.9380
26	Iron	Fe	55.847
27	Cobalt	Co	58.9332
28	Nickel	Ni	58.71
29	Copper	Cu	63.54
30	Zinc	Zn	65.37
31	Gallium	Ga	69.72
32	Germanium	Ge	72.59
33	Arsenic	As	74.9216
34	Selenium	Se	78.96
35	Bromine	Br	79.909
36	Krypton	Kr	83.80
37	Rubidium	Rb	85.47
38	Strontium	Sr	87.62
39	Yttrium	Y	88.905
40	Zirconium	Zr	91.22
41	Niobium	Nb	92.906
42	Molybdenum	Mo	95.94
43	Technetium	Tc	
44	Ruthenium	Ru	101.07
45	Rhodium	Rh	102.905
46	Palladium	Pd	106.4
47	Silver	Ag	107.870
48	Cadmium	Cd	112.40
49	Indium	In	114.82
50	Tin	Sn	118.69
51	Antimony	Sb	121.75
52	Tellurium	Te	127.60
53	Iodine	I	126.9044
54	Xenon	Xe	131.30
55	Cesium	Cs	132.905
56	Barium	Ba	137.34
57	Lanthanum	La	138.91

Atomic Number	Name	Symbol	Atomic Weight
58	Cerium	Ce	140.12
59	Praseodymium	Pr	140.907
60	Neodymium	Nd	144.24
61	Promethium	Pm
62	Samarium	Sm	150.35
63	Europium	Eu	151.96
64	Gadolinium	Gd	157.25
65	Terbium	Tb	158.924
66	Dysprosium	Dy	162.50
67	Holmium	Ho	164.930
68	Erbium	Er	167.26
69	Thulium	Tm	168.934
70	Ytterbium	Yb	173.04
71	Lutetium	Lu	174.97
72	Hafnium	Hf	178.49
73	Tantalum	Ta	180.948
74	Tungsten	W	183.85
75	Rhenium	Re	186.2
76	Osmium	Os	190.2
77	Iridium	Ir	192.2
78	Platinum	Pt	195.09
79	Gold	Au	196.967
80	Mercury	Hg	200.59
81	Thallium	Tl	204.37
82	Lead	Pb	207.19
83	Bismuth	Bi	208.980
84	Polonium	Po
85	Astatine	At
86	Radon	Rn	...
87	Francium	Fr
88	Radium	Ra
89	Actinium	Ac	...
90	Thorium	Th	232.038
91	Protactinium	Pa
92	Uranium	U	238.03
93	Neptunium	Np
94	Plutonium	Pu
95	Americium	Am
96	Curium	Cm	...
97	Berkellium	Bk	..
98	Californium	Cf
99	Einsteinium	Es
100	Fermium	Fm
101	Mendelevium	Md
102	Nobelium	No	...
103	Lawrencium	Lw	..

different. After traversing a circular path of 180 to 300 , the divergent particle paths are interrupted by catcher pockets, usually made of slots in graphite or copper, water-cooled "receivers." The isotopes are then chemically recovered from the receiver pockets.

Very small amounts of material can be separated in the calutron, since it separates the isotopes literally atom-by-atom. Nevertheless, ion currents up to one ampere can be maintained, allowing kilograms of material to be separated in a machine operating over a year's time. Virtually all of the isotopes of the elements have been separated in relatively high purity at Oak Ridge National Laboratory. The details on separated isotopes

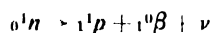
available and the procedures to be used in obtaining them are given in the ORNL Isotopes Catalog.

A number of gaseous elements, such as krypton, neon, and argon, are separated by thermal diffusion. Chemical exchange is used with such isotopes as hydrogen and nitrogen. Huge gaseous diffusion plants are used to enrich large quantities of ^{235}U . The electrolytic method is used for deuterium separation, and distillation has been used for the enrichment of mercury isotopes. More recently large high-speed centrifuges are being investigated for large-scale isotopic separations. Other than the well-known uses of ^{235}U and ^3H in large-scale nuclear work, the separated isotopes have been used primarily for fundamental scientific work, such as measurement of nuclear cross sections, but there is a growing utilization of isotopic materials in all fields of fundamental research and as target materials for radioisotope production.

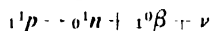
Radioisotopes. Some radioisotopes occur in nature, e.g., uranium, radium, and thorium—ordinarily accompanied by their radioactive daughters (decay products). Radioisotopes that occur in nature have half-lives greater than about 10^8 years or are the decay products of parent radioisotopes of such long-lived radioisotopes. These primordial radioisotopes were produced when the earth was formed and have not yet decayed away in the ensuing several billion years. Some shorter-lived radio-isotopes such as 5570-year ^{14}C and 12.46-year ^3H are formed by cosmic ray inter-actions with atmospheric nitrogen and hydrogen. Irene Curie observed and identified the first artificially induced radioactivity in 1934 by irradiating targets of aluminium, magnesium, and boron with alpha particles and noting that the targets continued to emit radiation after the α -source was removed. This discovery offered the first chemical proof of artificial transmutation. After the introduction of the cyclotron and other particle accelerators, many elements were bombarded with deuterons and protons to produce hundreds of new radioisotopes, including the well-known ^{131}I , ^{32}P , and ^{14}C . Large-scale production of radioisotopes, however, did not come about until nuclear reactors were available after World War II to supply enormous amounts of neutrons. Of the naturally occurring isotopes of the elements, roughly 280 are stable, and about 25 may be considered naturally radioactive. The number of artificially produced isotopes reached 200 in 1937, and with the nuclear reactor as a source of neutrons in World War II, about 450⁺ artificially radioactive isotopes were identified by 1944, and over 1500 by 1964. Each element has at least one radioactive isotope, and some have as many as 30.

Production of Radioisotopes. Radioisotopes are produced by disturbing a preferred neutron-proton ratio in the nuclei of elements. This is done by adding or removing neutrons, by adding or removing charged particles such as protons, or by a combination of both. Usually, a nuclear reactor is used as the source of neutrons; a cyclotron or other particle accelerator is used as the

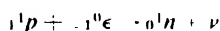
source of charged particles. The radionuclides formed by increasing the neutron-proton ratio generally decay (or transform) back to a stable condition by having a neutron transform to a proton, with the emission of a negative electron (beta particle, ${}_{-1}^0\beta$) and a neutrino (ν)—an almost undetectable uncharged particle of negligible mass. (see NUCLEAR REACTIONS).



For those radionuclides resulting from an increase in the proton-neutron ratio, the transformation again tends to reverse the cause for instability, and a proton in the nucleus is transformed into a neutron, with the emission of a positive beta particle (positron, ${}_{+1}^0\beta$).



In some cases, nuclei which are unstable because of an excess of protons (called neutron-deficient nuclei) will regain stability by capturing an orbital electron, ${}_{-1}^0e$:



Characteristic x-rays of the product element are emitted during the refilling of the orbital electron shells.

In the transformation processes described above, gamma (γ) radiation may or may not be emitted; these photons may undergo internal conversion whereby the transition between the two energy stages of a nucleus is not evidenced by the emission of a photon. Instead, the energy is imparted to an orbital electron, which is ejected from the atom.

For some nuclei, only gamma radiation is emitted for the de-excitation from a metastable or isomeric state. Such decay is termed isomeric transition (IT) and is characterized by no change in mass number or atomic number.

Many radioactive nuclides decay by two or more modes, so that ${}_{-1}^0\beta$, ${}_{+1}^0\beta$, γ , ${}_{-1}^0e$ transformations with associated emission and x-radiation are not uncommon. The *branching ratio* defines the relative amounts of each mode of decay.

For heavy nuclei ($Z \geq 82$), the transformation to a more stable configuration usually takes place by the emission of an alpha particle (α or ${}^4_2\text{He}$), sometimes accompanied by gamma emission.

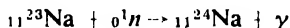
The bulk of the artificially produced radioisotopes are made by neutron reactions in the high-volume neutron fluxes available in NUCLEAR REACTORS. Neutrons, having no charge, can easily penetrate the coulombic barriers of the nucleus. The atomic nuclei of the elements vary in their ability to capture thermal neutrons (i.e., neutrons slowed down to ~ 0.04 eV) according to their *cross sections*, a value which expresses the probability of capture of a neutron of a certain speed as it passes near a nucleus (see CROSS SECTIONS AND STOPPING POWER). When target materials are placed in the reactor and subjected to a flux of neutrons (amount of neutrons traversing a unit

area per unit time), neutrons are captured in proportion to cross sections of the target element atoms present. Certain materials, such as aluminum and graphite, have such low neutron capture cross sections that few neutrons are captured; others, such as cadmium, have such high cross sections that a thin foil will absorb almost all the thermal neutrons impinging upon it.

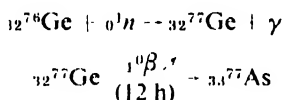
Radioisotopes are produced in a nuclear reactor by several different processes. Those processes that produce appreciable quantities of radioisotopes are described below.

(1) (n, γ) Process. In the (n, γ) process, which is most common, a neutron is captured by a target atom, and a gamma photon is emitted immediately. Since no change of the atomic number (charge on the nucleus) occurs, the element remains the same as the target material. The radioelement cannot be separated chemically unless a recoil collection is made, as in the Szilard-Chalmers process. The (n, γ) reaction is primarily a thermal-neutron (low-energy) reaction.

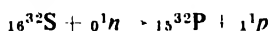
For example:



A radioisotope produced by this method sometimes decays by beta emission to a radioactive daughter with a different atomic number. The daughter can be separated chemically to obtain high specific activity* material. For example:



(2) (n, p) Process. In the (n, p) process, which requires neutrons of higher-than-thermal energies†, a neutron enters a target nucleus with sufficient energy to cause a proton to be released. The atomic number is reduced by 1, and the affected atom is transmuted into a different element, which can be separated chemically from the target material. Through the chemical separation, high-specific activity material can be obtained. For Example:



(3) (n, α) Process. The (n, α) process, like the (n, p) process, requires high-energy neutrons. In the (n, α) process, a neutron of high energy enters a target atom and causes an alpha particle to be emitted. The atomic number of the target atom is reduced by 2, and a chemical separation is possible. By means of chemical separation,

* *Specific activity* is the amount of radioisotope per unit weight of the total element and is usually expressed as curies or millicuries/gram.

† A few exceptions are found among reactions with the light nuclei in cases where the binding energy of a proton or particle is appreciably lower than that of a neutron: the reactions ${}^{10}\text{B}(n, p){}^{10}\text{Be}$, ${}^{14}\text{N}(n, p){}^{14}\text{C}$, ${}^{35}\text{Cl}(n, p){}^{35}\text{S}$, ${}^{10}\text{B}(n, \alpha){}^7\text{Li}$, and ${}^6\text{Li}(n, \alpha){}^3\text{H}$ occur with thermal neutrons.

high-specific-activity material can be obtained. For example:



(4) Fission. In the fission process, most of the fragments of uranium atoms which have undergone fission are radioactive atoms ranging from atomic number 30 through atomic number 64. They can be concentrated chemically for high specific activities, but since several isotopes of any one element are often produced, the isotopic purity will not necessarily be as high as that of radioisotopes produced by (n, p) and (n, α) reactions. The isotopic purity will depend somewhat upon the length of time that the uranium was exposed to neutrons and upon the elapsed time between removal from the reactor and the chemical separation.

The basic equation for radioisotope production is

$$A \xrightarrow{\sigma_A} B \xrightarrow[\sigma_B]{\lambda_B}$$

The target atom (A) captures neutrons to produce the product nuclide (B), which in turn is transformed by decay or further neutron capture. The effective cross sections (σ_A , σ_B) and the decay rate constant (λ_B) enable the rate and equilibrium values for the transformation to be calculated for any particular irradiation conditions. The exact solution for the differential equation describing these rate processes for the number of atoms (N) of product formed at time (t) in neutron flux ϕ is,

$$\lambda_B N_B = \frac{\lambda_B \phi \sigma_A N_A}{\lambda_B + \phi(\sigma_B - \sigma_A)} [e^{-\phi \sigma_A t} - e^{-(\phi \sigma_B + \lambda_B)t}]$$

In most cases, one can neglect the "burn-up" of the target atoms and the product radioisotope. In such cases, the above equation then reduces to:

$$\lambda_B N_B = N_{A0} \phi \sigma_A (1 - e^{-\lambda_B t})$$

Here N_{A0} refers to the number of original target atoms at time zero.

For irradiations of sufficient length ($t \gg T_{1/2}$), and again neglecting burn-up, the saturation factor ($1 - e^{-\lambda_B t}$) approaches 1 and the equation further reduces to

$$\lambda_B N_B = N_{A0} \phi \sigma_A$$

Isotope Processing. The techniques used for the processing of ultrahigh purity chemicals are required for isotope work in recovering stable isotopes, preparing target materials, and separating and purifying radioisotopes. Practically every technique from traditional wet chemistry to ion exchange and chromatography are utilized, often with high-purity radioisotopes, at very low concentration levels (e.g., micro-grams per liter). Sophisticated analytical methods (e.g., mass and radiation spectral analysis) are also required and make up a significant portion of the cost of isotope preparations.

A. F. RUPP
J. J. PINAJIAN

References

Nuclear and Radiochemistry

- Evans, R. D., "The Atomic Nucleus," New York, McGraw-Hill Book Co., Inc., 1955.
- Cork, J. M., "Radioactivity and Nuclear Physics," Third edition, Princeton, N. J., D. Van Nostrand, 1957.
- Friedlander, G., Kennedy, J. W., and Miller, J. M., "Nuclear and Radio-chemistry," second edition, New York, John Wiley & Sons, Inc., 1964.
- Oak Ridge National Laboratory, Catalog, Radio and Stable Isotopes, TID 18870, April 1963, available from Isotopes Development Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Stable Isotopes

- Baker, P. S., Bell, W. A., Jr., Davis, W. C., Ketron, C. V., Love, L. O., Martin, J. A., Olszewski, E. B., Prater, W. K., and Spainhour, K. A., "Production and Distribution of Electromagnetically Enriched Isotopes," *Proc. Intern. Conf. Peaceful Uses At. Energy*, 2nd, Geneva, 20, 245-50 (1958).
- Baker, P. S., "Stable Isotopes—Aid to Research," *Chem. Eng. News*, 37, 60-7 (1959).
- Baker, P. S., "Isotope Production," in "Encyclopaedic Dictionary of Physics," Section B/6, pp. 1-8, New York, Pergamon Press, 1961.
- Baker, P. S., "A Fifteen-Year Summary of Publications Involving the Uses of Electromagnetically Enriched Stable Isotopes," Oak Ridge National Laboratory Report, ORNL-3266, March 1963; available from Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U. S. Department of Commerce, Springfield, Virginia.
- Koch J. Ed., "Electromagnetic Isotope Separators and Applications of Electromagnetically Enriched Isotopes," Amsterdam, North-Holland Publishing Company, 1958.
- Smith, M. L., "Electromagnetically Enriched Isotopes and Mass Spectrometry," New York, Academic Press Inc., 1956.
- Wakerling R. K., and Guthrie, A., "Electromagnetic Separation of Isotopes in Commercial Quantities," TID 5217, United States Atomic Energy Commission, Division of Technical Information Extension, Oak Ridge, Tennessee, 1951, available from Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U. S. Department of Commerce, Springfield, Virginia.

Reactor Production of Radioisotopes

- Rupp, A. F., "Production and Separation of Radioisotopes," in Bradford, J. R., Ed., "Radioisotopes in Industry," pp. 168-189, New York, Reinhold Publishing Corp., 1953.
- Rupp, A. F., "Large Scale Production of Radioisotopes," *Proc. Inter. Conf. Peaceful Uses At. Energy*, 1st, Geneva, 14, 68-84 (1955).
- Rupp, A. F., "Radioisotope Production," in "McGraw-Hill Encyclopedia of Science and Technology," pp. 310-15, New York, McGraw-Hill Book Co., 1960.
- Rupp, A. F., "Production of Radioisotopes with Very High Specific Activity," *Proc. Fifth Japan Conf. on Radioisotopes*, Tokyo, (May 21-23, 1963).

- Rupp, A. F., "Reactor By-products," Reactor Technology Selected Reviews, 1964, USAEC Report, TID-8540, pp. 477-528.
- Rupp, A. F., and Binford, F. T., "Production of Radioisotopes," *J. Appl. Phys.* 24, 1069-81 (1953).
- Rupp, A. F., and Binford, F. T., "Production of Radioisotopes," in Etherington, H., Ed., "Nuclear Engineering Handbook," Section 14, pp. 26-37, New York, McGraw-Hill Book Co., 1958.
- Stehn, J. R., Goldberg, M. D., Magurno, B. A., and Wiener-Chasman, R., "Neutron Cross Sections," Vol. 1, z, to 1 to 20, Sigma Center, Brookhaven National Laboratory Report, BNL-325, and 2nd Ed., Supplement No. 2, May 1964.

- Aebersold, P. C., and Rupp, A. F., "Production of Short-lived Radioisotopes," in "Production and Use of Short-lived Radioisotopes from Reactors," Vol. 1, pp. 31-47, International Atomic Energy Vienna, 1962.

- Brookhaven National Laboratory, "Manual of Isotope Production Processes in Use at Brookhaven National Laboratory," BNL-864 (T-347), August 1964; available from Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U. S. Department of Commerce, Springfield, Virginia.

- Friend, C. W., and Jenkins, A. R., "Isotopes—A Program for Neutron Product Yield and Decay Calculations Using a Control Data 1604-A Computer," Oak Ridge National Laboratory Report, ORNL-3673, January 1965, available from Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U. S. Department of Commerce, Springfield, Virginia.

- Friend, C. W., and Knight, J. R., "ISOCRUNCH—Modification to the Crunch Program for the IBM-7090," Oak Ridge National Laboratory Report, ORNL-3789, January 1965; available from Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U. S. Department of Commerce, Springfield, Virginia.

- Knoll, Peter, "The Technology of Isotope Production," Part I, "Irradiation Technology," (Zentralinstitut für Kernphysik Dresden) ZfK-RCH-I, December 1961 (in German); available from International Atomic Energy Agency Library, Vienna.

- Oak Ridge National Laboratory, "ORNL Radioisotope Procedures Manual," ORNL-3633, June 1964, available from Clearinghouse for Federal Scientific and Technical Information, National Bureau of Standards, U. S. Department of Commerce, Springfield, Virginia.

- Roy, J. C., and Hawton, J. J., "Table of Estimated Cross Sections for (n, p) , (n, α) , and $(n, 2n)$ Reactions in a Fission Neutron Spectrum," Atomic Energy of Canada, Limited, Report, AECL-1181, December 1960.

Cross-references: ATOMIC PHYSICS, CROSS SECTIONS AND STOPPING POWER, ELECTRON MASS SPECTROMETRY, NEUTRON, NUCLEAR REACTIONS, NUCLEAR STRUCTURE, PERIODIC LAW AND PERIODIC TABLE, PROTON, RADIATION CHEMISTRY, RADIOACTIVITY.

K

KEPLER'S LAWS OF PLANETARY MOTION

The German astronomer and mathematician, Johannes Kepler (1571–1630), worked in the late 1500's with the Danish astronomer, Tycho Brahe, one of the most careful observers of astronomical motions of the pre-telescopic centuries. When Tycho Brahe died in 1601, Kepler inherited his data books to which he devoted several years of intensive work in an effort to bring orderly relations to light. Kepler was successful in deriving three experimental laws of planetary motion that led the way to the presently understood dynamics of the solar system. The first two laws were published in Prague in 1609, about the time when Galileo was first using his telescope to make significant discoveries about the planets and their moons; the third law did not appear until 1619.

Stated briefly, these laws are (see Fig. 1).

- (1) Each planet moves in an elliptical orbit with the sun at one focus of the ellipse.
- (2) The line from sun to any planet sweeps out equal areas of space in equal lengths of time.
- (3) The squares of the sidereal periods of the several planets are proportional to the cubes of their mean distances from the sun.

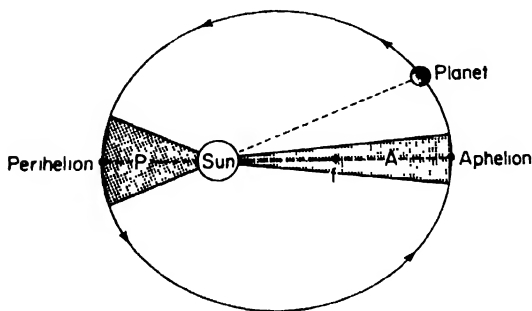


FIG. 1. Elliptic orbit of planet with Sun at one focus. Areas A and P are described in equal length of time.

An explanation of each of these laws follows, but first they should be seen against the background of the cosmology of Kepler's time, as they signify a distinct break with ancient cosmology and they were a marked extension of Copernican ideas. Until Kepler's time, it had been supposed

that the planets moved in circular orbits, that being the most perfect of curves and hence the path of the "perfect" celestial matter of which they were presumably composed. As observational accuracy increased, it became necessary to develop complex orbit by epicycles to preserve this idea of circular motion. The spectroscope would not show until after 1860 that heavenly matter and earthly matter were both composed of the same chemical elements. It is greatly to Kepler's credit that he established these three laws on such precise data and with such care, even though he himself did not know what they meant. In the ten years intervening between announcing laws (1) and (2) and publishing law (3), Kepler made innumerable attempts to find a relation between periods and distances in the solar system: some of these bordered on pure numerology and mysticism. Kepler had no adequate notion of force and hence he was prevented from continuing with his laws to their consequences, as Sir Isaac Newton did some five or six decades later.

Newton employed Kepler's laws to the full. The fact that planets move in ellipses with the sun at one focus he showed to be possible only if a force prevailed between sun and planet exactly proportional to the inverse square of the distance between them. This gave him confidence in the "inverse square law" (see GRAVITATION). The second law, Newton showed, was evidence of the great principle of conservation of angular momentum which Newton himself did so much to establish. Any object moving under a "central force," or a force directed toward a fixed point, will qualify for Kepler's second law, namely of sweeping out equal areas in equal times. However, although this law is not limited to inverse square forces, it shows, when coupled with the first law, that a planet such as the earth will move more rapidly about the sun at perihelion (closest to sun) than at aphelion (farthest from sun), as is observed. Finally, the third law falls into line as a necessary consequence if an inverse square law of force holds. One by-product of this law is the fact that the same force field is present from the primary body for each of the planets, diminishing with distance, but indicating that the force depends upon the *product* of the masses of sun and planet.

It is interesting to see how easily the third law of Kepler's may be "discovered" by the use of

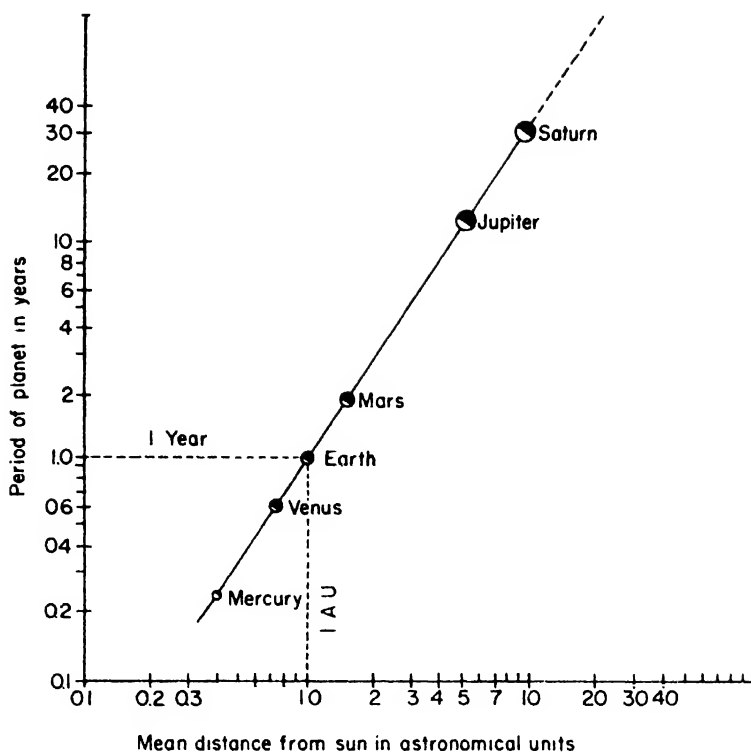


FIG. 2. Graph on log-log paper of relative mean distances of planets from sun, in Astronomical Units (where sun-earth distance is 1 A.U.), vs. period of planets about the sun in years. Straight-line graph with slope $3/2$ means constant ratio of R^3 to T^2 , which Kepler found in 1619. Planets beyond Saturn, unknown to Kepler, can be added.

techniques not available to Kepler. Kepler knew from Tycho's observations the *relative* distances from sun to any of the planets (out as far as Saturn); such relative distances can be found without knowing actual distances which were not found for another century. If one plots these relative distances against the planetary periods, using log-log graph paper, there results a fine straight line graph of points, one point for each planet, for which the slope is $3/2$ (Fig. 2). This shows at once that there is a constant ratio of R^3/T^2 ; or the ratio of the cube of R , mean distance from the sun, to the square of the sidereal period T , is the same ratio for all planets. Thus, in a few minutes, one may discover what it took Kepler ten years to find!

It was probably fortunate that Kepler made his most intensive study on the orbit of the planet Mars, inasmuch as that planet has the greatest eccentricity of orbit of any of the planets known in his time. Thus he could cast out circular orbits in favor of ellipses. (Only Pluto, discovered 300 years later has greater eccentricity.) One may, by relatively simple garden observation (non-telescopic), discover the eccentricity of the earth's own orbit by introducing a useful point of view. Ordinarily, astronomers talk about the place of the sun among the stars in its annual apparent pilgrimage along the ecliptic, or the plane of the

earth's orbit. However, it is not possible to see the stars when the sun is among them in the daytime, and it is therefore meaningful to ask "Where is the earth among the stars as seen from the sun?" This reversal of the point of view then means "What star (or position among the stars) is on the local meridian at true midnight, when the sun is on the anti-meridian beneath the observer's feet?" A vector drawn from sun to earth and thence to the stars would then be found to move eastward among the stars just as the earth proceeds eastward day by day around the sun. However, the rate of advance of this vector is not constant, although its *average* rate is only slightly less than one degree per day. Consequently, one would observe, by taking into account the equation of time (which reflects the variable apparent motion of the sun among the stars), that this "midnight vector" (see Fig. 3) does not progress uniformly: its rate of eastward motion is greatest near January 1 when it projects into the midst of the winter stars between Betelgeuse and Procyon; and it progresses least rapidly around July 1 when it projects into the region of the sky between Vega and Altair.

According to Kepler's second law, the product of angular velocity of the earth and the *square* of its radius vector from the sun should be constant, which means that in January when the earth is

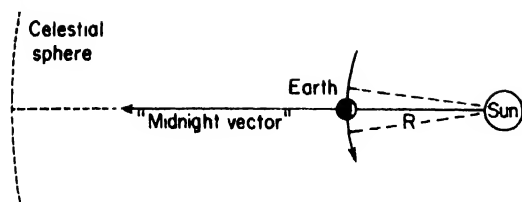


FIG. 3. The "place of the earth among the stars" is given by vector from sun to earth to celestial sphere (at midnight); the eastward sweep of this vector varies as R (the distance between sun and earth) varies, according to Kepler's second law; the angular speed of the vector multiplied by R^2 is constant.

nearest to the sun, its angular velocity as measured by this technique, should be greater by about 7 per cent than in July when the earth is some 3 per cent farther from the sun. Despite the fact that star days are all the same length, from the meridian crossing of any given star until its return to the meridian, the rate of progress of the true midnight meridian through the stars is not constant, but variable. It requires careful observation to measure the ratio of perihelion to aphelion distances in the manner suggested, but it can be done.

More simply, one may observe the application of Kepler's laws to the motion of the moon, for in one month one may discover that at certain times of the month when the moon is near perigee, or nearest the earth, it advances eastward among the stars at a greater rate than it advances about 13.5 days later when it is at apogee. Likewise, if one has any means for observing the apparent diameter of the moon from one end of its crescent to the other, one can also find that at perigee it subtends a greater angle by 11 per cent than at apogee. The variation in angular speed is quite pronounced inasmuch as the ratio of maximum to minimum distance is about 252/228 (in thousand miles) or 1.11. Conservation of momentum says that the product of angular speed and the *square* of the radius vector is constant, consistent with the constancy of areal velocity which depends on the square of the radius vector. Hence, the maximum angular speed divided by the minimum angular speed would be $(1.11)^2$ or about 1.24. The maximum speed exceeds the minimum by some 24 per cent, a quantity rather easily observed if one takes a little care and patience.

Kepler's laws are, of course, applicable to the motion of artificial or man-launched satellites

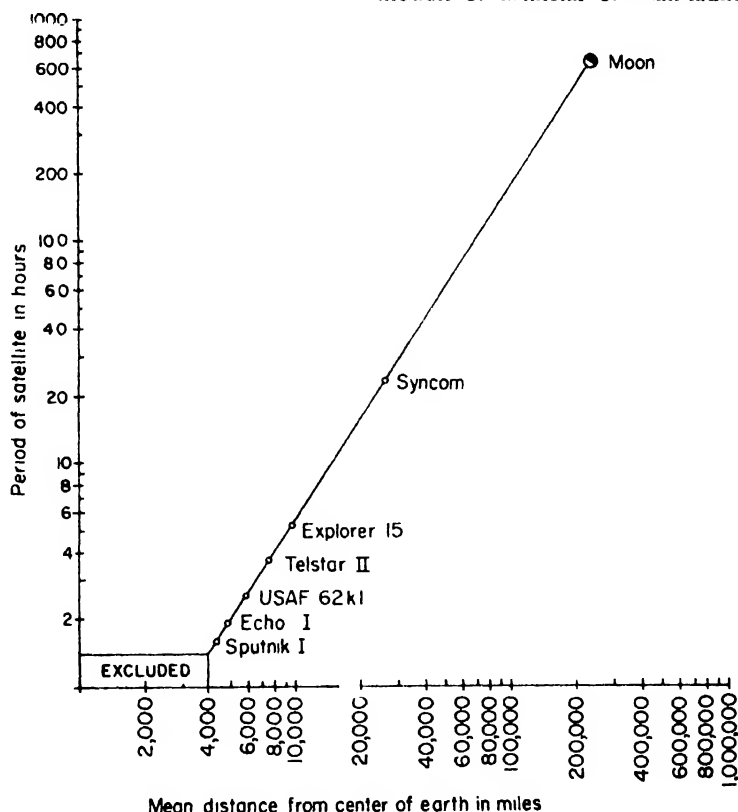


FIG. 4. Satellites of the earth - log log graph of mean distance from center of earth vs period of satellite, from Sputnik I (1957) to moon. This Figure is similar to Fig. 2 except that the earth rather than the sun controls the satellite periods. Again, the R^3/T^2 constant ratio shows up in the $3/2$ slope of the straight-line graph. Several representative satellites are listed.

circling the earth. One may draw a graph similar to that drawn for the planets, except that now the earth rather than the sun is the controlling body. Just as no satellite could possibly continue to orbit the sun if its mean distance were less than the radius of the sun, no satellite can continue to orbit the earth if its orbit intersects the earth. (In fact, close-by satellites are captured by the atmosphere and such satellites are likely to be burned up in the upper atmosphere as their kinetic energy is rapidly turned into heat.) It is interesting, however, to draw a straight line on our log-log graph from the point representing Sputnik I, the first artificial satellite launched in 1957 with its period of 96 minutes, to the moon with its period of 27.32 days. It will be found that other satellites lie as points on this line. The semi-major axis (or mean distance) of each satellite's orbit may be found from the addition of the earth's diameter to the usually published perigee and apogee distances, most commonly offered as minimum and maximum distances of the satellite above the earth's surface. Thus, for a satellite with minimum distance 1730 miles and maximum 2120 miles from earth's surface, we add 7950 miles (for the earth's diameter) to the sum of these figures and divide by 2 to obtain 5900 miles as the mean distance of the satellite from the center of the earth, the point toward which the gravitational forces are effectively directed. Armed with this 5900-mile mean distance, we may read the period of the satellite directly from the graph,

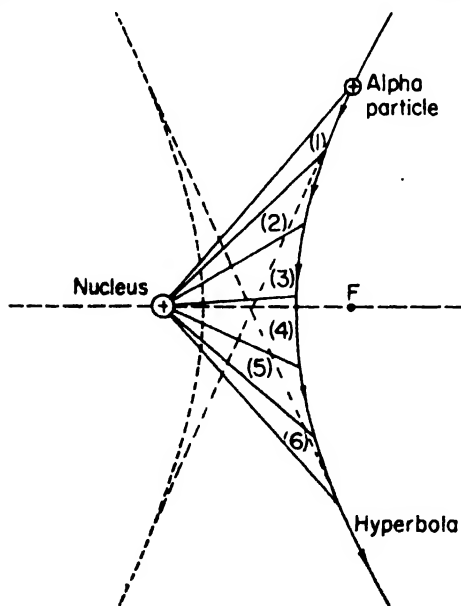


FIG. 5. Kepler's law of areas holds also for "central force" of repulsion between alpha particle and atomic nucleus; path of alpha particle is a hyperbola with nucleus in far focus. Areas 1, 2, 3, 4, 5, 6, etc., are equal and are described in equal lengths of time. A high-speed comet injected from outer space might, under gravity, describe the same kind of path, but with the sun at the near focus F.

namely about 153 minutes (2.55 hours). (This happens to correspond with USAF satellite 1962 K 1, launched on April 9, 1962 at such elevation that it is likely to continue to orbit for many years.)

It may be remarked in conclusion that Kepler's law of areas is applicable also to repulsive forces. The path of an atomic particle, such as a proton, in the field of an atomic nucleus is a hyperbola because of the strong repulsion between the two. In this case, the repulsive center is situated at the far focus of the hyperbola (see Fig. 5). Rutherford made full use of this idea in his brilliant discovery of the atomic nucleus (1911).

RICHARD M. SUTTON

Cross-references: ASTRODYNAMICS, ASTROMETRY, DYNAMICS, ROTATION - CIRCULAR MOTION, MECHANICS.

KERR EFFECTS

The Kerr effect is the occurrence of double refraction in a substance, when it is placed in an electric field. A more exact name is the *Kerr electro-optic* effect. It was first discovered for glass, in 1875, but it exists for liquids and gases also.

An ordinary dielectric medium has a single value for the index of refraction at any given wavelength of electromagnetic radiation, independent of polarization or direction of propagation of the radiation. In a uniaxial crystalline substance, there are two indices of refraction: the ordinary index which refers to radiation polarized with its electric vector perpendicular to the optic axis (a direction in the crystal), and the extraordinary index which refers to radiation with its electric vector parallel to the optic axis.

A substance which exhibits the Kerr effect is ordinarily singly refracting, but it becomes doubly refracting when an external electric field is applied. The substance behaves like a uniaxial crystal, with the optic axis parallel to the electric field. Liquids generally show larger effects than either solids or gases.

The difference in indices of refraction is proportional to the square of the applied electric field strength, and hence the Kerr effect is sometimes called the quadratic electro-optic effect. The expression for the optical path difference is given by

$$\Delta = (n_E - n_O)l = KE^2l/\lambda$$

where

- n_E extraordinary index
- n_O ordinary index
- l path length in the medium
- E electric field strength
- λ wavelength in vacuum
- K Kerr constant.

The constant is usually positive, decreasing with an increase in wavelength and with an increase in temperature. The effect is caused by either natural or induced optical anisotropy of the

individual molecules of the medium, and a lining up of the molecules by the applied electric field. The effect is particularly large for liquids made up of polar molecules with large anisotropies. Nitrobenzene, for example, gives an optical path difference of approximately $1/40$ wavelength in the visible region of the spectrum, for a field strength of 10000 volts/cm and a 1-cm path length. The Kerr effect can be very useful in helping to determine the structure of complex molecules.

Recent discoveries have shown the existence of an enormous quadratic electro-optic effect in certain crystals, notably perovskites in the paraelectric phase. The effect can be several thousand times larger than that observed for nitrobenzene and is independent of temperature. The theoretical explanation requires knowledge about both electronic and ionic states of the crystalline lattice constituents.

Kerr cells are light shutters constructed from substances showing large Kerr effects. One type of widely used cell contains nitrobenzene placed between capacitor plates or electrodes. The cell itself is situated between a crossed polarizer and analyzer combination. In the absence of an electric field, no light is transmitted. When the electric field is applied, its strength is chosen so that the cell becomes a half-wave plate (the optical path difference is one-half wavelength). The plane of polarization of the light is therefore rotated by 90° , and the light is transmitted. An important property of the cell is its extinction ratio, or ratio of the light transmitted when the cell is closed to that transmitted when it is open. A cell filled with extremely pure nitrobenzene can give a ratio as low as 10^{-5} to 1. Cells of the perovskites are closer to 10^{-2} to 1. Another important property of the Kerr cell is the frequency with which it can be turned on and off. Frequencies as high as 10^{10} /sec are attainable in some instances. The inductance and capacitance of the cell, as well as relaxation effects in the medium, must be given considerable attention in the design of a cell. Also, the transit time of the light across the cell must be much less than the period of the applied field, in order for the cell to be effective.

A second effect is known as the *Kerr Magneto-optic* effect. It is the change in polarization of light reflected from a polished metal pole of a magnet. Linearly polarized light incident normally, is reflected as elliptically polarized light. This effect has not as yet proved useful in a practical device, nor for determination of fundamental properties of substances, and therefore has not been investigated as thoroughly as the electro-optic effect.

SUMNER P. DAVIS

References

Electro-optic Effect:

- Szivessy, G., "Handbuch der Physik," Vol 21, p. 724, Berlin, Springer, 1929 (theory and experimental detail).
Geusic, J. E., Kurtz, S. K., Van Uitert, L. G., and

Wemple, S. H., *Appl. Phys. Letters*, 4, 141 (1964) (properties of perovskites).

Kerr Cells.

Zarem, A. M., Marshall, F. R., and Poole, F. L., *Elec. Eng.*, 68, 282 (1949).

Magneto-optic Effect:

Schütz, W., *Wien-Hurms Handbuch der Experimental Physik*, 16, 319 (1936).

Cross-references: POLARIZED LIGHT, REFRACTION.

KINETIC THEORY

The kinetic theory is a branch of THEORETICAL PHYSICS developed in the nineteenth century to explain and calculate the properties of fluids. It is most useful for studying the physical properties of gases, but it can also be applied to liquids, electrons in metals, and neutrons passing through solids. The word "kinetic" means "pertaining to motion," in this case the motion of molecules or subatomic particles.

Historical Development. The first attempt to develop a kinetic theory of gases was made in 1738 by the Swiss mathematician Daniel Bernoulli. Bernoulli began with the idea that matter consists of tiny atoms moving about rapidly in all directions, which the Greek philosopher Democritus had presented, but was unable to prove. Bernoulli showed that the collisions of atoms against the walls of a container would produce a pressure which would be inversely proportional to the total volume of the container; he assumed that the space occupied by the atoms themselves is negligible compared to the total volume of the container and that the rest of the space is empty. He also found that the pressure would be directly proportional to the kinetic energy of motion of the atoms if the velocities of the atoms are changed while the volume is kept fixed. (The kinetic energy of an atom is half its mass multiplied by the square of its velocity.)

The British scientist Robert Boyle had already shown in 1662 that the pressure of air varies inversely as its volume if the temperature is held constant (see GAS LAWS). Thus Bernoulli's theory was able to explain a well-known fundamental property of air and other gases. It was also known that the pressure of a gas confined in a fixed volume increases with temperature. However, it was not until about 1800 that there was enough experimental evidence, and an accurate enough temperature scale, for Gay-Lussac (French) and others to establish a quantitative relation between pressure and temperature. This relation can be expressed by saying that pressure is proportional to the temperature measured from "absolute zero" (though it was not until later in the nineteenth century that the idea of absolute zero temperature was generally accepted). According to the kinetic theory, the absolute temperature is proportional to the kinetic energy of motion of molecules in a gas.

The kinetic theory was proposed again in the first half of the nineteenth century by two British

scientists, John Herapath and J. J. Waterston. Neither of them was familiar with Bernoulli's theory, which had not made much impression on the world of science. Waterston obtained one important new result, which is now known as the "equipartition theorem": in a mixture of two or more different gases at the same temperature, the average kinetic energy of each kind of molecule will be the same. This means that heavy molecules will tend to move more slowly than light molecules, since when the mass of a molecule is greater, its velocity must be less in order to keep the kinetic energy the same.

In 1859, the German physicist Rudolf Clausius showed how the kinetic theory could be used to explain the rate of mixing of two gases and the rate of heat conduction. His work was extended by James Clerk Maxwell (British), who calculated the viscosity coefficient by the kinetic theory. He found that theoretically the viscosity of a gas should be the same at different pressures, and should increase with temperature. This seemed to go against common sense, but later experiments by Maxwell himself and other physicists showed that the theory is correct. Soon afterward several scientists, starting with Josef Loschmidt (Austrian) in 1865, used the kinetic theory to calculate the diameter of an atom. At this time it began to appear that the atom is something that really exists in nature, since it can be measured, weighed, and counted, and is not merely a philosophical speculation. By then the atomic theory had already been accepted in chemistry as a basis for explaining chemical reactions, but it was the kinetic theory of gases that established the place of atoms in physics.

Starting from the foundations laid by Clausius and Maxwell, Ludwig Boltzman (Austrian) and J. Willard Gibbs (American) worked out systematic methods for calculating all the properties of gases from kinetic theory (see STATISTICAL MECHANICS). Sydney Chapman (British) and David Enskog (Swedish) completed the theory, insofar as it pertains to the transport properties (diffusion, viscosity, and heat conduction) of gases at ordinary densities, although there are still some unsolved problems in the area of high-density gases and liquids, on the one hand, and rarefied (very low density) gases on the other. In the course of working out this theory, Chapman and Enskog discovered that it should be possible to separate the components of a mixture of a gas by making one side of the container hotter than the other. This effect—known as "thermal diffusion"—was soon afterwards established experimentally by Chapman and Dootson (British), thus confirming the prediction based on kinetic theory. (Thermal diffusion was used as one of the methods of separating isotopes of uranium during the development of the atomic bomb in World War II).

Although the kinetic theory was founded on the principles of classical Newtonian mechanics and led to some incorrect results because those principles are not valid on the molecular level, it is now generally agreed that the kinetic theory

is valid for calculating the statistical properties of large numbers of molecules, provided that the properties of the individual molecules themselves are determined experimentally, or from the quantum theory. It is only when one tries to apply the kinetic theory to matter in extreme conditions (very low temperatures or very high densities) that he must take account of quantum-mechanical modifications of the statistical method itself (see QUANTUM THEORY and STATISTICAL MECHANICS).

Main Features of the Theory. By assuming that the major part of the heat energy of a gas consists of kinetic energy of motion of the molecules, one finds that the average velocity of a molecule is several hundred meters per second under ordinary conditions. However, it is a fact of common observation that gases do not actually move as a whole at such speeds; a gas will eventually spread throughout any container in which it is placed, but it may be several seconds or minutes, for example, before chlorine gas generated at one end of a large laboratory is noticed at the other end. According to the kinetic theory, the reason for the relative slowness of gaseous diffusion, in contrast to the high average velocities of individual molecules, is that a molecule can travel on the average only a very short distance (its "mean free path") before it collides with another molecule and changes its direction of motion. In particular, at atmospheric pressure, if the molecular diameter is assumed to be about 0.0000001 cm (which is approximately true for most molecules), the mean free path would be approximately 0.0001 cm. At the same time, the average distance between neighboring molecules would be somewhat more than 0.000001 cm, so that the fraction of the total volume occupied by the molecules themselves is less than 1 part in 1000. The average molecular velocity in air at 15°C is about 460 m/sec, so that a molecule will have about 4 600 000 collisions per second.

To simplify their calculations, Maxwell and Boltzmann made the following assumptions:

(1) Instead of trying to compute the exact path followed by every molecule, they assumed that, because of the enormous frequency of collisions, the velocities and positions of molecules in a gas are distributed at random over all possible values consistent with the known physical state of the gas. For example, it is assumed that the average total velocity is known, as is the temperature (which fixes the mean square velocity). If variations of temperature and density from one place to another can be ignored, then the molecular velocities can be described by a statistical distribution—the "Maxwell distribution"—which is similar to the normal "bell-shaped curve" or law of errors in statistics.

It should be noted that the effect of a large number of collisions is *not* to make all the velocities equal, but rather to produce a wide range of velocities from zero up to very large values—though the probability of large deviations from the average is quite small. The existence of this

"spread" of molecular velocities has been verified directly by various experiments.

(2) The diameter of a molecule is so small, compared to the average distance between molecules, that simultaneous collisions of three or more molecules may be ignored. The validity of this assumption for low-density gases makes it possible to develop a very accurate theory of gas properties, since these properties can be related to the interactions of molecules taken two at a time, and the mathematical description of such two-particle interactions is relatively simple. The corresponding kinetic theory of dense gases and liquids, on the other hand, involves the solution of difficult many-particle problems, and results can be obtained only with the help of very drastic approximations (see MANY-BODY PROBLEM).

(3) In treating the collision of two molecules, one supposes that there is no correlation between their velocities before the collision. In the elementary kinetic theory, one assumes that each molecule has the average velocity characteristic of the region of the gas in which it has most recently undergone a collision. It thus "forgets" its past history every time it collides with another molecule. This is obviously not strictly true for each individual molecule, but it is a useful approximation for dealing with average properties of large numbers of molecules.

In modern physics research, it is usual to distinguish between "equilibrium theory" and "transport theory," both of which grew out of the elementary kinetic theory. Equilibrium theory is described in the article on STATISTICAL MECHANICS (see also EQUILIBRIUM); it is used to study such properties as the heat capacity (amount of heat needed to raise the temperature by a certain amount) and the compressibility (change in volume produced by a small change in pressure). Equilibrium theory also tries to explain the existence of phase transitions, such as the condensation of gases to liquids, or the appearance of magnetic ordering in solids. From the theoretical viewpoint, the calculation of equilibrium properties is simpler than that of transport properties such as viscosity, because one merely averages over all the possible states of the system (i.e., over all possible combinations of velocities and positions of the molecules) without having to worry about how one state follows another in time. TRANSPORT THEORY involves a detailed analysis of molecular collisions in order to determine how changes in the state of the system are related to external forces or nonuniform conditions imposed on it.

One of the most fruitful techniques in transport theory is the use of "Boltzman's equation." This equation describes how the velocity distribution changes as a result of external forces and collisions between molecules. Unfortunately the equation is rather difficult to solve, because the term that expresses the effect of collisions on the velocity distribution is an integral over the values of the (unknown) velocity distribution itself for two colliding molecules. In order to

calculate the transport properties it is necessary to resort to tedious computations with infinite series, except for certain artificial force laws (such as repulsive forces inversely proportional to the fifth power of the distance between two molecules) for which the integral can be simplified. In most cases, the results (as worked out by Chapman and Enskog) do not differ greatly from the ones obtained from the approximate elementary theory. However, the important phenomenon of thermal diffusion was discovered only because of a theoretical prediction by Enskog and Chapman; since the existence of this effect had not even been suggested by the earlier theories, this discovery must be regarded as one of the triumphs of mathematical analysis.

Irreversibility. As indicated above, the kinetic theory assumes that the velocity of a molecule may depend on the conditions in the region where it has just suffered a collision, but is otherwise random—in other words, independent of its previous history. This assumption permits one to use the methods of probability theory even though, in classical mechanics, the actual motions of the molecules are regarded as completely determined by their initial configurations. As long as one uses the theory only to calculate properties of a gas that can actually be measured during a relatively short time, the assumption of randomness leads to no serious errors. However, it introduces an element of irreversibility which is inconsistent with the reversibility of the laws of classical mechanics. (A reversible process is one that can go equally well forwards or backwards, in contrast to an irreversible process, like scrambling an egg, which cannot be undone without a great expenditure of energy.) The irreversible aspect of the kinetic theory is shown most clearly by Boltzmann's "*H*-theorem," which has led to a considerable amount of controversy about the foundations of kinetic theory. Boltzmann showed in 1872 that a certain quantity, later called *H*, which depends on the velocity distribution, must always decrease with time, unless the velocity distribution is Maxwell's distribution, in which case *H* remains constant. In the latter case, which corresponds to the equilibrium state, *H* is proportional to the negative of the entropy (see article on THERMODYNAMICS). Thus the *H*-theorem provides a molecular interpretation of the second law of thermodynamics or, in particular, the principle that the entropy of an isolated system must always increase or remain constant. Irreversible processes are those in which entropy increases. The entropy itself can be regarded as a measure of the degree of randomness or disorder of the gas, although it must be recognized that disorder really means just our own lack of knowledge about the details of molecular configurations. The equilibrium state represents the maximum possible disorder; the *H*-theorem implies that a gas which is initially in a nonequilibrium (partly ordered) state will eventually reach equilibrium and then stay there forever if it is not disturbed.

If the long-term consequences of the *H*-theorem

were applicable to all matter in the universe, one might expect that the universe would eventually "run down"; although the total energy might always remain the same, no useful work could be done with this energy because all matter would be at the same temperature (see THERMODYNAMICS). This final state has been called the "heat death" of the universe.

The contradiction between the *H*-theorem and the laws of classical mechanics is shown by two famous criticisms of the kinetic theory, the "reversibility paradox" and the "recurrence paradox." The first paradox is based on the fact that Newton's laws of motion are unchanged if one reverses the time direction, so that it would seem to be impossible to deduce from these equations a theorem that predicts irreversible behavior. In deriving the *H*-theorem, Boltzmann assumed that the velocities of two molecules are uncorrelated *before* they collide. Obviously they must be correlated *after* they have collided, so the theorem will not be correct if the time direction is reversed. The second paradox is based on a theorem of Henri Poincaré (French): if a mechanical system is enclosed in a finite volume, then after a sufficiently long time it will return as closely as one likes to its initial state. Hence *H* must return to its original value; if it has decreased during some period of time, it must increase during some other period. The time between successive recurrences of the same state for the molecules in 1 cc of air is much longer than the present age of the universe, so one does not have to worry about recurrences in any actual experiment. In spite of the two paradoxes, *H* is always found to decrease with time in actual experiments.

Other applications of the kinetic theory are discussed in the articles on AERODYNAMICS, BOLTZMANN'S DISTRIBUTION LAW, ELECTRICAL CONDUCTIVITY, LIQUID STATE, NUCLEAR REACTORS, and PLASMA.

STEPHEN G. BRUSH

References

- For an elementary introduction, see Cowling, T. G., "Molecules in Motion," London, Hutchinson, 1950; reprinted by Harper Torchbooks, New York, 1960.
- Diffusion and thermal diffusion: Furry, W. H., "On the Elementary Explanation of Diffusion Phenomena in Gases," *Am. J. Physics*, **16**, 63 (1948).
- Comprehensive treatments of the elementary theory and many applications: Loeb, L. B., "Kinetic Theory of Gases," New York, McGraw-Hill Book Co., 1934; Kennard, E. H., "Kinetic Theory of Gases, with an Introduction to Statistical Mechanics," New York, McGraw-Hill Book Co., 1938.
- Theory of the Boltzmann equation: Chapman, S., and Cowling, T. G., "The Mathematical Theory of Non-Uniform Gases," Cambridge, Cambridge University Press, 1952; Grad, H., "Principles of the Kinetic Theory of Gases," in "Handbuch der Physik," Vol. 12, p. 205, Berlin, Springer, 1958.
- Reprints of original papers, with historical introduction: Brush, S. G., Ed., "Kinetic Theory," New York, Pergamon Press, 1965.
- Calculations of properties of fluids: Hirschfelder, J. O., Curtiss, C. F., and Bird, R. B., "Molecular Theory of Gases and Liquids," New York, John Wiley & Sons, 1954.

L

LASER

"Laser" is an acronym for *l*(ight) *a*(mplification by) *s*(timulated) *e*(mission of) *r*(adiation). This device is identical in theory of operation to the MASER except that it operates at frequencies in the optical region of the electromagnetic spectrum, rather than in the microwave. Laser operation has been demonstrated at wavelengths from 3000 to over 1 000 000 Å or from 0.3 to 100 μ . By common usage, these devices are all called lasers, although more descriptive terminology utilizes ultraviolet maser, optical maser, infrared maser, etc. Although the original microwave maser offers an extremely stable frequency source, its main use is as an amplifier with extremely low noise output. In contrast, the main significance of the laser is its ability to produce a single frequency at high intensity in the optical region, a feat heretofore impossible at these frequencies. Not only may the output be a single monochromatic wave, but the wave may be coherent, or in phase, over the whole surface of the radiator. In this mode of operation, the laser is actually an oscillator whose output depends upon the selective amplification of one of the single frequency modes of the resonant cavity containing the active laser medium.

Following the development of the microwave maser, Schawlow and Townes in 1958 proposed that optical maser action could be obtained by placing an active medium in an optical cavity. The medium would be a gas or solid which was excited electrically or by light in such a manner that any optical wave present would be amplified as it moved through the material. The cavity was proposed to be a Fabry-Perot resonator—two plane parallel reflecting plates with a small transmission through which the radiation might escape. Upon excitation of the material, light will be emitted with a band of frequencies determined by the particular material. In addition, the direction of emission will be nominally random. In the presence of the cavity, some of the waves will escape after several back and forth reflections from the parallel plates, "walking off" the edge of the reflectors. Those waves which travel normal to the walls will remain in the cavity and be amplified provided they reinforce each other after each round-trip reflection at the two surfaces. This reinforcement or resonance is only satisfied if the spacing of the plates is an integral multiple of one-half the wavelength in the medium. Thus, after a

short time, only that frequency which satisfies the resonant condition and those waves traveling normal to the reflector will build up to an appreciable intensity. The resultant light which is partially transmitted through one of the reflectors will thus be a single frequency or several discrete frequencies if there is more than one cavity resonance within the band of frequencies emitted by the laser material. In addition, the wave front will be in phase across the surface of the reflector since waves striking the surface at normal incidence are amplified most strongly. The resultant beam will then be diffraction limited, i.e., the beam will spread by an angle in radians given approximately by the ratio of the wavelength to the diameter of the beam. In actual practice, single-mode operation is obtained only under special conditions. Generally, several frequency modes are present due to the multiple resonances of the cavity and numerous "off-axis" modes are found which correspond to resonant waves which travel at small angles from the normal to reflectors. These waves "walk off" so slowly that they still are amplified appreciably. Refinements of the simple cavity proposed by Schawlow and Townes consist of concave reflectors which decrease the diffraction losses or several parallel reflectors which limit the oscillation to a frequency common to each pair in the set.

The key to successful laser operation is of course the active medium which amplifies the wave. Qualitatively, a material which fluoresces or exhibits luminescence is an obvious candidate. In fluorescence, electrons are excited to an upper-energy state by short-wavelength light such as ultraviolet, while luminescence is produced by passing an electron current through the medium, such as in a gaseous discharge. In either process, stimulated emission can occur only if more electrons are produced in the upper-energy state than in the lower or terminal state for the radiating transition. In this case, an incident photon will stimulate further transitions and amplification will result. If the final state were more heavily populated, then the photon would cause more upward or absorbing transitions and the net effect would be absorption.

The first optical maser was demonstrated by Maiman of Hughes Research Laboratories in 1960 using ruby, which is single-crystal aluminum oxide "doped" with chromium impurities. By applying semitransparent reflective coatings on

the ends of a rod about 2 inches long, he made the cavity and the crystal an integral unit. Then, exposure to an intense exciting light from a xenon flashtube was found to invert the population between the red-emitting level and the ground or lowest-energy state of the electrons. The result was a burst of intense red light, emanating in a beam through the end reflectors. This was the first and is still one of the most powerful lasers. Advances in the art since that time have resulted in energies per pulse of the order of 1000 joules or watt-seconds. Peak powers are as high as 500 000 kW in short pulses of the order of 10^{-8} seconds. Because of "off-axis" modes and multiple resonances, the output is not a single-frequency, single plane-wave mode, but generally consists of the order of 100 separate modes. The beam is still quite narrow, being the order of 1 milliradian or 0.05 degrees. As a comparison with conventional light sources, the energy radiated from 1 cm² of the brightest flash lamp is less than 10 kW and is distributed over the entire visible spectrum. In addition, the radiation is incoherent and is spread out uniformly in all angles from the source. Thus, the directivity and spectral purity of the laser source are many orders of magnitude superior to that of an incandescent source. The ruby laser suffers from a low efficiency, about 1 per cent, and except with elaborate cooling systems, only operates on a pulsed basis. Other crystalline or glass systems with impurity ions have been developed, which yield wavelengths from the ultraviolet to approximately 3μ wavelength in the infrared. None of these are as powerful as ruby with the exception of glass doped with neodymium, or rare earth ion, which is comparable in power and radiates at 1.06μ in the near infrared, but with a broad distribution of spectral lines.

Historically, the next development came in 1961 when Javan, Bennet and Herriott demonstrated laser action in a gaseous discharge of helium and neon. Again, the parallel-plate reflector cavity was used but this time with a spacing of several feet. Later, concave mirrors were used to decrease the loss of energy out the sides of the cavity. This device operates continuously and delivers power at levels up to one watt. Pulsing the gas discharge yields peak powers as high as 100 watts. The first laser radiated at 1.15μ in the infrared, while further development with different gases has yielded outputs from the ultraviolet to 33μ or 0.33 mm in the far infrared. In contrast to the ruby laser, the gaseous laser beam may be diffraction limited and the frequency is pure, i.e., oscillation may be limited to one mode. By careful design, the frequency may be stabilized to a few thousand cycles per second or approximately one part in 10^{13} . The gas discharge is excited either by short-wave rf power or by a dc current supplied by internal electrodes in the discharge tube. Here the population inversion is produced by electron and ion collisions with the radiating atom. Again, as in ruby, the operating efficiency is low, less than 1 per cent.

The third main type of laser utilizes a solid

material, in this case a semiconductor. Here the electron current flowing across a junction between *p*- and *n*-type material produces extra electrons in the conduction band. These radiate upon making a transition back to the valence band or lower-energy states. If the junction current is large enough, there will be more electrons near the edge of the conduction band than there are at the edge of the valence band and a population inversion may occur. To utilize this effect, the semiconductor crystal is polished with two parallel faces perpendicular to the junction plane. The amplified waves may then propagate along the plane of the junction and are reflected back and forth at the surfaces. The gain in the material is high enough so that the reflection at the semiconductor-air interface is sufficient to produce oscillation without special reflective coatings. The first such device used gallium arsenide and radiated at 8400\AA or just beyond the visible region in the infrared. This laser was developed by groups at General Electric, International Business Machines, and Lincoln Laboratory in 1962. The efficiency is high, about 40 per cent, and the power source is low-voltage direct current. One shortcoming is the requirement of liquid nitrogen cooling (77 K) to maintain power output and efficiency. Powers as high as 3 watts continuous have been produced. The cavity in this case is extremely small, the reflector spacing being less than a millimeter. As a result, it is fairly easy to limit the oscillation to one frequency mode although small irregularities in the junction prevent coherence over the full width of the narrow radiating junction strip. The compactness and efficiency of the semiconductor laser make it particularly attractive for systems use. Wavelengths as long as 12μ and as short as 6300\AA have been generated using different semiconductors such as indium arsenide, indium phosphide, indium antimonide, or alloys such as gallium arsenide-phosphide.

Lasers have also been operated in liquid media, utilizing rare earth ions in such organic hosts as chelates. The most significant liquid laser utilizes a different principal than those above, depending upon stimulated Raman scattering. Raman laser action was discovered by Woodbury in 1962 using a ruby laser and nitrobenzene. Here the laser excites the nitrobenzene, which in turn shows amplification at a frequency displaced from the ruby line by the vibrational frequency of the molecule. There is no true inverted population in this case. The incident photon is scattered by the molecule which absorbs an amount of energy determined by its vibrational energy. The molecule is left in an excited state and the scattered photon is frequency shifted by the energy loss. This process may be stimulated, since the rate at which the scattered photons are produced is proportional to the number of photons already present in the cavity at the scattering wavelength. As in the normal stimulated emission case, the frequency and phase of the output wave are identical with the wave which stimulates the scattering. The Raman laser normally operates

using the Stokes line, or the wavelength corresponding to the loss of one vibrational quantum. Other modes of operation utilize the second or third Stokes lines corresponding to double or triple vibrational absorptions. Similarly, higher-order effects in the medium may produce a series of anti-Stokes lines which correspond to vibrational energy being added to the initial energy of the photons from the driving laser. The wavelength range of Raman lasers using different liquids is from the visible to the near infrared.

The high instantaneous powers quoted for ruby are obtained by using the "Q-switched" mode of laser operation. This technique, due to Hellwarth and McClung, uses a cavity resonator whose reflectivity or "Q" may be controlled externally. The laser, usually ruby, is first excited by the flash lamp while the cavity is in a state of low reflectivity and thus low feedback. As a result, the inverted population reaches an extreme value before oscillation occurs. At the peak of inversion, the reflectivity is "switched on," and the resultant high reflectivity produces an intense burst of energy which almost completely depopulates the high-energy states in a time of the order of 10^{-8} seconds. The switching is accomplished either by a Kerr electro-optic shutter in the cavity or by rotating one of the mirrors so that it is lined up parallel with the opposite reflector at the optimum time during the flash lamp pulse.

R. H. KINGSTON

References

1. Schawlow, A. L., "Advances in Optical Masers," *Sci. Am.*, **209**, 34-35 (July 1963).
2. Lengyel, B. A., "Lasers," New York, John Wiley and Sons, Inc., 1962.
3. Lengyel, B. A., "Optical Masers," New York, Polytechnic Press, 1963.
4. Grivet and Bloembergen, Eds., "Quantum Electronics III, I, II," New York, Columbia University Press, 1964.

Cross-references: COHERENCE, LIGHT, MASER, OPTICAL PUMPING, RAMAN EFFECT AND RAMAN SPECTROSCOPY.

LENS

A lens is any element that focuses light to form images. Many lenses are found in nature. Ice crystals, waves on the surface of water, and all the eyes of humans and animals are examples of lenses. These lenses have one or more curved surfaces and are made of a transparent material. Manufactured lenses are usually made out of glass or crystal material. The simplest lens consists of two ground and polished spherical surfaces. A line connecting the centers of the two spheres is called the optical axis of the lens. The lens is edged to form a cylindrical surface centered on the optical axis. The spherical surfaces may be convex or concave, resulting in lenses which are positive refracting or negative

refracting. A positive lens collects the light from a distant object and focuses it to a real image. The negative lens disperses the light and causes it to diverge from a virtual image. Positive and negative elements are used in combinations to form optical lens systems. The optical axes of each of the lens elements usually coincide to form centered optical systems. Most optical systems are designed to be centered optical systems, but in manufacture the centering is seldom perfect, so the system will have various degrees of defective performance.

Spherical surfaces are usually used in optical systems because of stringent requirements on the manufacture of optical elements. In order to perform properly, a given surface in an optical system often has to coincide with the prescribed surface to within a few millionths of an inch. Such extreme tolerances can be achieved on spherical surfaces because spheres may be ground and polished with self-correcting techniques.

A few lenses have been made using non-spherical surfaces, but they usually have rotational symmetry around the optical axis. These surfaces are called rotationally symmetric aspheric surfaces. Aspheric surfaces of this type are difficult to generate so they are used infrequently.

Some lenses are made with cylindrical and toric surfaces. Spectacle lenses often have surfaces of this type. It is practical to use aspherics in spectacle lenses because the beam of light entering a person's eye is small in diameter. The performance requirements are therefore not great. Cylindrical or toric surfaces are seldom used in telescopes, or microscopes of high performance.

There are many types of glass used in optical lenses. Some glasses are more dispersive (see REFRACTION) than others. By combining positive and negative lenses of different glass it is possible to correct for chromatic aberrations.

Theory of the Lens. Most of the performance of a lens or lens system may be understood by considering that light travels as rays in straight lines until it encounters a change of index of refraction. The light is then refracted according to Snell's law (see REFRACTION). Light is emitted from a point source of light in the object as a diverging beam of rays. A lens is able to collect these rays and refocus them to an image point (see Fig. 1).

With analytical geometry, one may derive equations for calculating the path of any ray as it passes through the optical system. The procedure is called ray tracing. The mathematical equations used to trace rays are long and complicated, and have to be computed with many significant figures. Prior to the use of modern digital computers, the design and analysis of lens systems was a long tedious job. An average lens design required many months of calculation. Today most of these calculations are done on large computers, and few people need to be concerned about being able to ray trace.

Paraxial Rays. Paraxial rays pass through the center portion of the lens and the assumption is made that the object points are close to the optical axis. The ray-tracing equations for

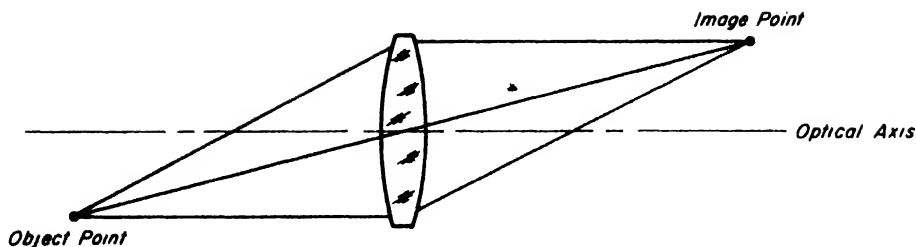


FIG. 1. A diagram showing how a lens collects diverging rays and focuses them at an image point.

paraxial rays are simple, and by using the paraxial approximation, many useful theorems for lens performance may be worked out.

In the paraxial region, any optical system may be described by locating six cardinal points along the optical axis. Once these cardinal points are known, the position and size of the image of any object may be computed from the following formulas (see Fig. 2).

$$m = \frac{y_k}{y_0} = -\frac{z'}{f} = -\frac{f'}{z} \quad (1)$$

$$zz' = ff' \quad \text{and} \quad \frac{f}{s} + \frac{f'}{s'} = 1 \quad (2)$$

$$f/n_0 = f'/n_k \quad (3)$$

$$P_1N_1 = P_2N_2 \quad (4)$$

$$P_1P_2 = N_1N_2 \quad (5)$$

$$F_1N_1 = f' \quad (6)$$

$$N_2F_2 = f \quad (7)$$

P_1 and P_2 are called the first and second principal points. N_1 and N_2 are called the first and second nodal points. F_1 and F_2 are the first and second focal points. f and f' are the front and back focal lengths. n_0 and n_k are the indices of refraction in the object and image space.

The following terms are commonly used in connection with lenses:

Field of View. The field of view usually refers to the half angle subtended by the object as seen from the first principal point P_1 . For example, it would be $\tan^{-1}(y_0/S)$ in Fig. 2. When specified for a lens, it usually refers to the maximum size

of object which may be imaged by the lens. Optical designers tend to describe the field of view by its half angle, as shown in Fig. 2. Marketing firms often refer to the full field which is twice the half angle. If not clearly stated, confusion over this term may result.

Relative Aperture. This refers to the half angle of the cone of rays converging to the axial image point. The sine of this angle is often called the numerical aperture and is written NA . If NA is large, the lens collects a large cone of light and focuses it on the image. Another way to describe this NA is to use the term f -number. The f -number of a lens and the NA of a lens are related by the following equation

$$f\text{-number} = \frac{0.5}{NA}$$

Aperture Stop. The aperture stop in a lens system is a diaphragm which determines the NA of the lens.

Lens Aberrations. Lens designers attempt to combine elements and glass types to reduce the lens aberrations. All points in the object should be imaged as points in the image and should be located at or very near the position predicted by the paraxial rays. In lenses of large relative aperture and field of view, there are usually several residual aberrations that designers are unable to eliminate. There are the following types of aberrations:

- (1) Spherical
- (2) Coma
- (3) Astigmatic
- (4) Field curvature
- (5) Distortion
- (6) Axial chromatic aberrations
- (7) Lateral chromatic aberrations.

These aberrations are corrected by using combinations of positive and negative elements. There are two general principles one may use as guide lines in correcting optical systems. (1) A closely spaced positive and negative lens with the aperture stop in contact may be corrected for spherical aberration, coma, axial and lateral chromatic aberration and distortion. (2) It is necessary to use positive and negative lenses with appreciable air space between them to correct field curvature and astigmatism.

There are many conflicting requirements in lens systems. Lenses of large relative aperture usually are designed to cover small fields of view.

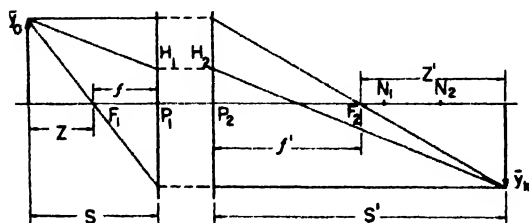


FIG. 2. Diagram showing the location of the six cardinal points in a lens system.

A large field of view normally dictates a small relative aperture. Wide fields and large aperture may be realized by using many elements or by compromising some of the image quality. For example, lenses of 140° total field working at $f/2$ are available, but they are complex and have large amounts of distortion. Periscopes allow one to look through a long pipe and see a wide field at high magnification, but there is always some residual chromatic aberration and field curvature left in the design.

Lenses are usually designed for specific applications, and the designer has made a careful balance between conflicting requirements. It is seldom that a lens designed for one application will perform optically in another. In the past there was a tendency to misuse lenses, because it was difficult to design a new system. Today it is easy to obtain a new design. It is still expensive to build a new prototype design. Optical shop practice has not yet been appreciably affected by our modern technology. This is partly because there has not been sufficient economic pressure to stimulate the investment in equipment, but it is also because the tolerances are still beyond the capabilities of modern automated machine tools. The next few years should see many changes in the optical industry, and it is expected that new systems will be much more readily attainable. With this improvement in optical shop capability, we will see many improvements in optical system capabilities because optical systems will be designed for specific tasks instead of for general use.

Lens Testing. A designer specifies a set of curvatures, thicknesses and optical glasses for a nominal design. The work shop makes the lens to these specifications within certain tolerances. The final lens must then be tested to make sure that the over-all performance is as expected. It is rare indeed that a lens performs exactly as computed. It is then necessary to determine if the difference is negligible and, if not, what to do about it. First, one tests the complete assembled lens and attempts to predict the performance. By studying the defective image, it may be possible to locate the sources of error. The tests consist of studying the light distribution in the image of a point source. This is done on a lens bench or testing interferometer. Sometimes it is possible to locate the source of error by testing the over-all system, but usually the lens system is disassembled and all the lens surfaces, spaces and centering are checked separately.

Lens Types. There are several optical systems which may be classified as types. There is considerable overlap between the types, but there is some value in the following classification.

Microscope Objectives. Microscope objectives are used to magnify small objects. They are usually used with a microscope eyepiece. Microscope objectives range in focal lengths from 2 to 48 mm and are used at magnifications ranging from 1000 to $5\times$. The high-power objectives are made up of many small elements. Some of the lenses are only a few millimeters in diameter. The lens

making and mounting procedures for such small lenses are quite different from larger elements.

Telescope Objectives. Telescope objectives are used to view distant objects. Telescope objectives have a wide variety of focal lengths and diameters. They are usually corrected precisely for spherical aberration, coma and axial chromatic aberration. Since telescope objectives cover small fields of view, astigmatism and field curvature usually are not corrected. Telescope lenses of large diameter (20 inches or more) become afflicted with chromatic aberration which cannot be corrected, so many of the large telescopes used by astronomers are mirrors instead of lenses.

Telescope objectives are used in binoculars, opera glasses, surveying instruments, gunsights, and many laboratory instruments.

Periscopes. Periscopes are used to enable one to look through a long tube. The submarine periscope is one well-known example, but there are many other types used in industrial and medical instruments. Gastrosopes and cystoscopes are examples of periscopes used in medical instruments. Periscopes are made up of a train of telescope objectives and eyepieces.

Camera Objectives. By far the largest class of optical lens systems would be classified as camera lenses. They are used to record images on films as in common landscape cameras, but today they are also used with many other types of image recording systems such as television image tubes, electrostatic plates, etc. The distinguishing features of camera lenses are wide field and large aperture. Usually the image is located on a flat image plane. Camera lenses range in complexity from a single meniscus lens to systems with more than ten elements. Camera lenses cover such a wide range of uses that one could claim all lenses to be a form of camera lens. For example, a long focal length lens used on a 35-mm camera may actually be very similar to a telescope objective.

Eyepieces. Eyepieces are quite clearly a distinctive class of optical lens system. Eyepieces are designed to match the sensitivity and physical requirements of the human eye. For example, the eye is sensitive in the visual part of the spectrum so eyepieces are designed for this range of wavelengths. An eyepiece must also be located with its aperture stop in an external position so that the observer's eye may be located within it. This requirement imposes serious limitations on eyepieces, and they are seldom useful in any other applications.

Magnifiers are essentially eyepieces except they are designed to view opaque material while an eyepiece is designed to view an aerial image formed by an objective.

Condensers. Condensers are used to collect and focus large amounts of light. They are found in projectors and substages of microscopes. A searchlight mirror is a form of a condenser. The numerical aperture of a condenser is usually very large, and for many applications, the image-forming properties are not important. Condensers are often made, therefore, with low-quality surfaces. Some condenser lenses are

molded. Condensers are usually placed close to an intense light source which heats and cracks the lenses if made of glass. Condensers are often made out of quartz because of its ability to withstand heat. Aspheric condenser lenses are common in condensers. With some of the modern high-intensity light sources, it is necessary to correct for the image errors in order to obtain uniform illumination.

ROBERT E. HOPKINS

References

- Hardy, A. C., and Perrin, F. H., "Principles of Optics," New York, McGraw-Hill Book Co., 1932.
 Greenleaf, Allen, "Photographic Optics," New York, The Macmillan Co., 1950.
 Conrady, A. E., "Applied Optics," London, Oxford University Press, 1929.

Cross-references: ABERRATIONS; MICROSCOPE; OPTICAL INSTRUMENTS; OPTICS, GEOMETRICAL; REFRACTION.

LIGHT

Light is a form of electromagnetic energy. It has a physical character similar to that of radiowaves. In order that the human eye may get the sensory perception of light, the electromagnetic waves entering the pupil should have a wavelength λ between 4000 and 7000 Å ($1 \text{ Å} = 10^{-8} \text{ cm}$). The wavelength of a wave is inversely proportional to its frequency ν . The product of the two quantities equals the velocity of propagation. For light in vacuum one has $c = \nu\lambda = 3 \times 10^{10} \text{ cm/sec}$. The frequency of light waves is therefore almost a billion times higher, their wavelength a billion times shorter, than the waves of standard radio broadcast bands. The perception of color depends on the distribution of the electromagnetic energy over the visible wavelengths. White light is a superposition of waves at many frequencies. It can be decomposed into its monochromatic spectral components by a prism or other spectral apparatus. The violet end of the spectrum is near 4000 Å, the red end near 7000 Å. Whereas light in its narrow definition should be confined to this relatively narrow portion of the electromagnetic spectrum, it is customary to extend the definition to the ultraviolet and infrared portion of the spectrum. One sometimes speaks loosely of ultraviolet and infrared "light," although electromagnetic waves at these frequencies are not detected by the eye. The human mind and hands have, however, devised a large variety of instruments by which such invisible radiation can be detected and measured. Photographic plates can be made sensitive to x-rays, with a wavelength shorter than the ultraviolet, or to the much longer wavelengths of the infrared. Geiger counters can detect electromagnetic radiation of very short wavelengths (γ -rays and x-rays). Photoelectric cells are sensitive in the ultra-violet and visible portion of the spectrum. Photoconductivity can

be used to detect infrared radiation. At still longer wavelengths, the microwaves and radiowaves are detected by diode detectors in appropriately arranged microwave and radioreceivers. All these types of radiation can also be converted into heat by absorption in a blackbody, i.e., a material that can absorb radiation at all wavelengths. The radiation can be felt as heat, if it is absorbed by the human skin.

The study of the human eye as a detector of light is the task of *physiological optics*. The impression of light is not necessarily always connected with the simultaneous presence of electromagnetic energy at the retina. We see "stars" from a heavy mechanical blow in the dark. The impression of light is retained for about 0.1 second after the light source is shut off. This fact is made use of in the movies to create the impression of motion by a series of still images. The eye is a detector with a relatively long response time. Photoelectric cells can react more than a million times faster. Color vision is also subject to physiological peculiarities which are quite complex (see COLOR and VISION AND THE EYE).

The property of light which is most immediately accessible to observation is its propagation along straight lines (shadows). The law of REFLECTION also dates from Egyptian antiquity. If light rays pass from one medium to another, their direction is changed according to the law of REFRACTION. If the light in medium 1 propagates with a velocity v_1 and makes an angle θ_1 with the normal to the boundary between media 1 and 2, the direction θ_2 in medium 2, with a velocity of propagation v_2 is given by Snell's law, $\sin \theta_1 / \sin \theta_2 = v_1 / v_2 = n$. The constant n is called the relative index of refraction of medium 2 with respect to medium 1. These three laws are the basis of *geometrical optics*. This branch of the science of light describes the paths of light rays, the formation of images by mirrors and lenses, the action of telescopes, microscopes, prisms and other optical instruments.

The wave character of light becomes apparent by more refined observations. The phenomena of diffraction, interference and polarization are the subject of *physical optics*. Diffraction describes how waves are bent around obstacles. They represent corrections to and deviations from the laws of geometrical optics. These effects become pronounced only when the material has a characteristic dimension comparable to the wavelength of the wave. When light waves reach the same point along different paths, the resulting intensity may be smaller than that produced by each individual wave separately. The relative phases of the waves may be such that they interfere destructively, when the arrival of one wave with maximum positive deflection coincides with that of another wave with maximum negative deflection. Observations of light in crystals of calcite (iceland spar) first showed that there are two different modes of vibrations for each direction of propagation. These are called the two transverse modes of polarization.

All phenomena of geometrical and physical optics are described consistently by Maxwell's

equations of electromagnetic theory. Optical phenomena are, therefore, closely related to other electric and magnetic phenomena. Around 1900 the prevailing opinion was that the wave character of light was unambiguously established and the nature of light well understood.

There was, however, a mathematical difficulty with the intensity of radiation at ultraviolet and higher frequencies. The photoelectric effect could also be interpreted only by considering light to have a quality of particles. The number of electrons emitted from a photosensitive surface is proportional to the intensity of the light. The energy of the individual electrons is, however, determined by the light frequency. This led to the postulate of light quanta with energy $h\nu$, where h is Planck's constant. This duality in nature, in which "wave-like" and "particle-like" properties are combined, is described without internal contradiction by quantum mechanics. The combined "particle and wave" character of light is revealed by the combination of properties of the light sources, the electromagnetic field describing the light waves, and the detectors.

The study of the interaction of light waves with matter in the sources and in the detectors is the subject of SPECTROSCOPY. This is a wide field which encompasses atomic and molecular spectroscopy, parts of solid-state physics and photochemistry. The quantum theory was largely developed on the basis of spectroscopic data. A light quantum is emitted by an excited atom, molecule or other material system, when an electron in such a particle makes a transition or "quantum jump" from a state with higher energy to a state with lower energy. The energy difference between these states is equal to the quantum energy $h\nu$. Similarly, the absorption of light quanta is accompanied by an electronic transition from a state with a lower energy to a state with an energy higher by an amount $h\nu$. In this manner, the frequencies of spectral lines are characteristic for the electronic energy levels in each material. The frequency of the light may be said to correspond to the frequencies of the vibrating charges or oscillators, which are represented by the electrons.

Light sources are thus bodies with a sizeable population of electrons in excited states. This may be accomplished by raising the temperature of the material. The most important source of light is the sun. The moon and other planets are visible only because they reflect sunlight, just as all other objects on the earth which we can see by daylight, but not at night. The sun is a star. In stars, the temperature is maintained at a very high temperature by nuclear reactions.

Man-made light sources range from the primitive fire, candles, and oil and kerosene lamps to the electric light bulb, fluorescent gas-discharge tubes, arcs, etc. In early sources, the material particles of smoke or wick were heated by the chemical reaction of oxidation or burning; in the incandescent electric lamps, a wire is heated to a very high temperature by an electric current. There are so many energy levels in these luminous

solid materials or gases at high pressures that the emitted light is white and contains essentially all frequencies. The higher the temperature, the more radiation is emitted and the higher the average frequency of radiation. It should be realized that most of the energy is emitted as invisible (infrared) radiation, even in the best incandescent lamps. Hot gases in flames may also emit sharp spectral lines characteristic of the atoms occurring in the flame. The yellow color which arises when kitchen salt is sprinkled in a flame is due to the characteristic yellow spectral line of sodium atoms.

In gas discharge tubes, atoms or molecules are excited by collisions with electrons in the ionized gas. The energy is provided by the generator which provides the voltage necessary to maintain the discharge current. An arc is a discharge in air or in a high-pressure vapor. Mercury and sodium discharges are used for street lighting. Fluorescent tubes for office and home lighting use a gas discharge with a substantial amount of ultraviolet components. This ultraviolet light excites electrons in fluorescent centers on the walls of the tube. The electrons drop immediately from the highly excited state to an intermediate state with a lower energy. From this state they finally drop down to the original ground energy level with the emission of visible light. Gas discharges at relatively low pressure may serve as spectroscopic sources to study the emission spectra of atoms, ions and molecules. From the relationship between the energy levels and the frequency of radiation, it follows that a material, when heated, can emit precisely those frequencies which it absorbs, when it is in the lowest energy level at low temperature.

All these light sources are incoherent in the sense that there is no phase relationship between the light waves emitted by the different atoms in the source. This is quite different from the property of the usual sources of electromagnetic radiation at lower frequencies. In the oscillator tubes of radio- or microwave transmitters, all electrons move and vibrate in step with each other. The analogy between light and low-frequency electromagnetic radiation raises the question, "Can coherent light sources be constructed?" Recently such coherent light sources have been developed. They are characterized by the emission of a highly directional, highly monochromatic light beam of high intensity. They are called LASERS because they are based on light amplification by stimulated emission of radiation. In the conventional sources, all light is emitted spontaneously. In lasers, the original spontaneously emitted light forces the other excited atoms to emit their radiation in step, or coherently. If stimulated emission thus dominates the spontaneous emission, a laser results. This requires a high concentration of excited atoms and a sufficient feed back mechanism of light by mirrors. In its simplest form a laser consists of a gas discharge in a tube of suitably chosen dimensions and gas pressure between a set of parallel mirrors. Because the atoms in the laser source all act constructively in step, these sources provide a more efficient means to transmit light energy.

The high light intensities available in focused laser beams have led to the development of the branch of nonlinear optics. The optical properties of materials are different at high intensities, because the electronic oscillators are driven so hard that anharmonic properties become evident. A typical effect is the harmonic generation of light in which red laser light is converted into ultraviolet light at exactly twice the frequency, when the high-intensity beam traverses a suitable crystal such as quartz. It should be possible to duplicate at light frequencies all nonlinear effects known from the field of radio communications, such as modulation, demodulation, frequency mixing, etc. It is no longer correct to say that the propagation of a light wave is independent of the presence of other light waves. At high intensities, there is a noticeable interaction between light waves of different frequencies.

The combination of the laws of quantum mechanics and electromagnetic theory gives a consistent description of the generation, propagation and detection of light. Since these same laws also describe many other properties of matter such as electronic structure, chemical binding, electricity and magnetism, etc., it may be said that the nature of light is well understood. In this context, it is not necessary and not even desirable to pose the question, "What is it, precisely, that vibrates in a light wave in vacuum?" The electromagnetic fields acquire meaning only through their relationships with detectors and sources. Human knowledge or understanding is here used in the operational sense that a relatively simple framework of physical concepts and mathematical relationships exists, which gives an accurate description of the wide variety of optical phenomena at present accessible to observation or verification in experimental situations. The following references will introduce the reader to the vast literature of optics and spectroscopy.

N. BLOEMBERGEN

References

- Whittaker, E. T., "A History of the Theories of Aether and Electricity," Vols. I and II, London, Nelson & Sons, 1952.
 Born, M., and Wolf, E., "Principles of Optics," London and New York, Pergamon Press, 1959.
 Schiff, L. I., "Quantum Mechanics," New York, McGraw-Hill Book Co., 1955.
 Ditchburn, M., "Light," New York, Interscience Publishers, 1963.
 Minnaert, M. G. J., "Light and Color in the Open Air," Ann Arbor, Mich., Dover Publications, 1953.

Cross-references: ELECTROMAGNETIC THEORY; INFRARED RADIATION; INTERFERENCE AND INTERFEROMETRY; LASER; OPTICS, GEOMETRICAL; OPTICS, PHYSICAL; PHOTOCONDUCTIVITY; PHOTOELECTRICITY; QUANTUM THEORY; REFLECTION; REFRACTION; SPECTROSCOPY; ULTRAVIOLET RADIATION; VISION AND THE EYE.

LIGHT SCATTERING

When a beam of light falls on a particle, part of this incident beam is diverted from its original path; that part which is diverted and not absorbed is *scattered*.

Light scattering is a very familiar phenomenon. The colors of visible objects (other than light sources) are determined by the wavelengths which they scatter most effectively. Scattering by small particles was first studied experimentally in great detail by Tyndall (1869) in connection with the blue of the sky and has become known as *Tyndall scattering*.

Classical physics is appropriate for the description of most light-scattering phenomena. Thus, light scattering is explained in terms of the forces exerted by the electromagnetic field on the electronic charges which all matter contains. The oscillating electromagnetic field of the incident light exerts a periodic force on each electronic charge, causing it to execute harmonic motion at the light-wave frequency. It is the fact that an oscillating charge radiates in all directions (except along the line of its motion) which accounts for the scattering. The intensity of the radiation scattered from a particle will be large in directions for which the radiation from the individual elements of the particle interferes constructively, and small in directions in which it interferes destructively.

For particles comparable in size to the light wavelength, the amount of energy scattered as well as the angular distribution of the intensity and polarization of the scattered light are influenced by the distribution of induced oscillating charge within each scatterer. Any correlation which may exist between the positions of the scatterers also affects the extent to which the radiation interferes constructively or destructively to make up the resultant scattered field. Thus, in principle, light scattering provides a tool for the investigation of the number, size, shape, internal structure and orientation of particles and their mutual interactions.^{1,3}

The problem of relating the light scattering to these properties and vice versa has proved too difficult for solution in general. This is not surprising, for it would be necessary to solve Maxwell's equations with the proper boundary conditions for each particle. Many important cases have been solved, however, subject to certain approximations. Widest success has been achieved for *single scattering*, i.e., for particles sufficiently dispersed that radiation scattered by any one particle can be considered to escape from the medium without being further scattered by other particles. Multiple scattering is relatively difficult to treat and is usually avoided when possible.^{2,4}

In many cases, light scattering is related to the composition and structure of the medium in a way similar to x-ray scattering. The criterion which must be satisfied is that the electromagnetic field within the scatterers should be closely approximated by the unperturbed incident field,

just as in the x-ray case. Light scattering under this approximation is widely known as *Rayleigh-Gans* scattering. It is applicable if the phase shift for radiation passing through a particle is not too different from the phase shift which would occur for radiation passing through the same distance in the surrounding medium.^{1,3} When the Rayleigh-Gans approximation is valid, the angular intensity distribution of the scattered light is related, as in x-ray scattering, to a "form factor" which describes structure of the individual scatterers and to a "radial density function" or correlation function which describes the order in their spatial arrangement.^{2,3}

If the particles are less than about 1/10th of the light wavelength, and if their index of refraction is near to that of their surroundings, only the induced electric dipole radiation is important. Lord Rayleigh (1871) explained Tyndall's principal results in terms of the intensity and polarization of the induced electric dipole radiation. This type of scattering has since become known as *Rayleigh scattering*.

Rayleigh scattering is of particular importance. If the particles are dispersed at random (molecules of an ideal gas or widely dispersed macromolecules in an essentially homogeneous solution), the individual particles may be regarded as independent sources. In this event, the total scattered intensity is merely the sum of the intensities scattered by the individual particles. The special case of isotropic particles and unpolarized light is both simple and illuminating. The Rayleigh formula is

$$\frac{\text{Intensity of scattered light}}{\text{Intensity of incident light}} = \frac{8\pi^4 N \alpha^2 (1 + \cos^2 \theta)}{\lambda^4 r^2}$$

where N is the number of particles, α is their polarizability, θ is the angle of scattering, λ is the wavelength, and r is the distance from the scattering system to the point of observation (where $r \gg$ any relevant dimension of the scattering system). Thus, Rayleigh scattering from independent particles is proportional to the number of particles and is quite insensitive to their shapes. When the total mass of scatterers is known, it provides a tool for the measurement of molecular weight.

As we consider larger particles which begin to violate the criterion that their dimensions be very small compared with the wavelength, the Rayleigh formula breaks down. This breakdown first appears at large scattering angles, where the destructive interference is first significant, and quickly spreads to moderate and small angles. Nevertheless, for scattering angles sufficiently near zero, the Rayleigh formula retains validity since for zero scattering angle the radiation from all volume elements within a particle is "in phase" regardless of the particle size. Thus, the Rayleigh formula is actually useful, when properly applied, over an extremely wide range of molecular weights (10^2 to 10^7).¹⁻³

For scattering from dense media such as liquids, it is important to recognize that the individual molecules can not ordinarily be treated

as independent scatterers. Perhaps the most direct formulation of the problem is in terms of the radial density function mentioned earlier. Often, however, this function is not of immediate interest and one would prefer to relate the light scattering directly to the thermodynamic properties of the medium. For Rayleigh scattering, this may be accomplished in a direct way through an ingenious approach due to Smoluchowski (1908) and Einstein (1910).^{2,3} This approach takes advantage of the fact that for molecules small compared with the wavelength and for intermolecular forces extending over distances small compared with the wavelength, the scattered field may be regarded as made up of radiation from elements of volume small enough that each element may be considered an electric dipole source and yet large enough that the elements can be considered to be independent of each other. If the index of refraction of every element were identical, the solution would be homogeneous and no scattering would result. But the index of refraction of an element will fluctuate according to the number of molecules it contains. The total scattering is found to be proportional to the mean square fluctuation in index of refraction which is related to the thermodynamic properties of the solution through free energy.

Up to this point, the frequency of the scattered light has been regarded as identical to that of the incident light. Actually, line broadening will occur due to a number of mechanisms. The line structure has been measured when highly monochromatic laser light is used and has been interpreted in terms of particle size.⁵ In general, the spectrum of the scattered radiation is found also to have relatively weak lines (or bands) which are absent from the incident light. They were first studied in detail by Raman (1928), and are known as the *Raman spectrum*.⁶ It is perhaps the only light-scattering phenomenon which must be explained by quantum theory.

R. W. HART

References

1. van de Hulst, H. C., "Light Scattering by Small Particles," New York, John Wiley & Sons, Inc., 1957.
2. Kerker, M., Ed., "Electromagnetic Scattering," New York, The Macmillan Co., 1963.
3. Oster, G., "The Scattering of Light and Its Applications to Chemistry," *Chem. Rev.*, **43**, 319-365 (1948).
4. Beckmann, P., and Spizzichino, A., "The Scattering of Electromagnetic Waves from Rough Surfaces," New York, The Macmillan Co., 1963.
5. Cummins, H., Knable, N., Yeh, Y., "Observation of Diffusion Broadening of Rayleigh Scattered Light," *Phys. Rev. Letters*, **12**, 150-153 (1964).
6. Cleveland, Forest F., "Raman Spectroscopy," in Clark, G. L., Ed., "The Encyclopedia of Spectroscopy," pp. 675-681, New York, Reinhold Publishing Corp., 1960.

Cross-references: LIGHT; OPTICS, PHYSICAL; RAMAN EFFECT AND RAMAN SPECTROSCOPY.

LIQUEFACTION OF GASES

The liquefaction of all readily available gases has become a routine operation in industrial technology. Prominent among the reasons for converting a gas to a liquid are the net saving in cost of storing or of transporting a normally gaseous material in liquid form, the convenience and flexibility of providing very low temperature refrigeration to a multiplicity of sites of modest or intermittent consumption in the form of a low-boiling liquid, and the efficiency attainable in the separation of the components of a gaseous mixture by the partial liquefaction of the mixture, or its total liquefaction followed by rectification.

The transoceanic shipment of liquefied natural gas, the commercial distribution of liquid helium to scientific laboratories, and the production of pure oxygen and pure nitrogen from air are representative examples of the first, second and third reasons, respectively. The first and last examples currently operate on scales such that thousands of tons of liquid are produced daily.

To produce a cold liquid product from gaseous raw material at ambient temperature requires a heat pumping operation. Thermodynamic analysis gives the (unattainable) irreducible minimum work which must be expended in the heat pump, operating in an environment at temperature T_0 , to convert a unit mass of warm gas to liquid to be

$$W_{\min} = (H_{\text{liquid}} - H_{\text{gas}}) - T_0(S_{\text{liquid}} - S_{\text{gas}})$$

where H_{liquid} and H_{gas} are the enthalpies and S_{liquid} and S_{gas} are the entropies per unit mass of liquid product and gaseous raw material, respectively. These thermodynamically reversible works of liquefaction are listed for various of the "permanent" gases in Table I, which assumes that the starting material is gas at one atmosphere pressure and 300°K. In large-scale, practical operations, the actual work requirement will

range from ~3 times the minimum for a gas such as methane to ~15 times the minimum for helium.

TABLE I. MINIMUM WORK OF LIQUEFACTION OF VARIOUS GASES

Substance	Boiling Point (°K)	Work Required (kW-hr/lb)
Methane	111.7	.145
Oxygen	90.2	.080
Nitrogen	77.3	.096
Hydrogen	20.4	1.48
Helium	4.22	0.86

To achieve the minimum thermodynamic work requirement for cooling and liquefying a stream of gas, an infinite sequence of perfectly efficient refrigerators operating at successively lower temperatures ranging from ambient to the boiling point of the material would be required. Various approximations to this theoretical ideal have been developed.

Cascade Process. If the critical temperature of the gas which is to be liquefied lies well *above* the boiling point of some second fluid, whose critical temperature in turn lies above the boiling point of yet another fluid, and so on to some fluid which is condensable at ambient temperatures, then one can replace the infinite sequence of refrigerators of the thermodynamic ideal with this discrete series of liquid cooling baths.

The raw material is compressed to the pressure necessary to condense it at the temperature of the final refrigerant bath. The resulting liquid is expanded through a throttle valve, and the vapor which boils off in the throttling is recycled to conserve the refrigeration it represents. Such an arrangement is shown schematically in Fig. 1.

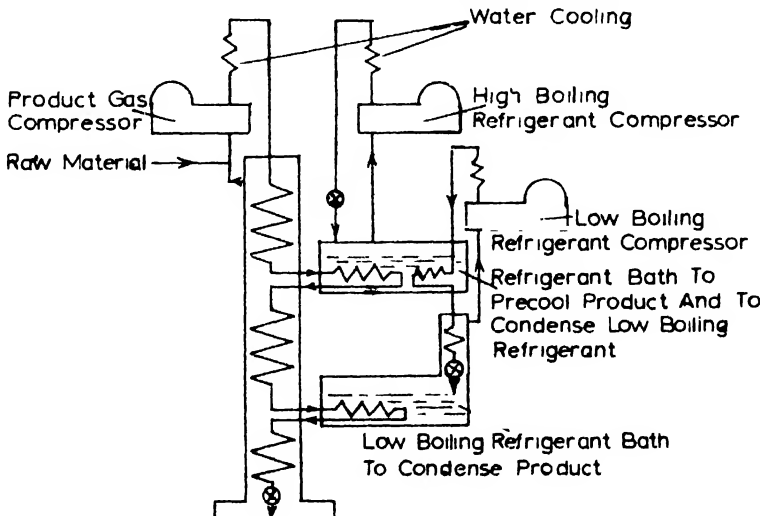


FIG. 1. Schematic arrangement for two-stage cascade.

The penalty in increased work over the thermodynamic minimum arises from the small number of steps in the sequence, with the attendant irreversible exchange of heat between the process gas and the much colder baths; from the throttling losses for the product liquid; and from the imperfect efficiency of *any* real refrigerator or compressor. In the simple liquefaction of any gas for which a cascade of refrigerants can be found, the economic performance of a cascade compares very favorably with any other process for truly large-scale operations.

For small-scale systems, the operational complexity and the equipment cost of a cascade are prohibitive.

Linde and Claude Processes. A stream of cold gas, flowing countercurrent to the process stream in a heat exchanger which establishes perfect thermal equilibrium between the two streams at every point along their path can substitute perfectly for the infinite sequence of refrigerators in the theoretical ideal system. The problem is just to produce the stream of cold gas (let alone to produce it with perfect efficiency) and to produce a refrigerator to extract the heat of vaporization from the product material at its boiling point.

Application of the first law of thermodynamics to the system shown in Fig. 2, consisting of a

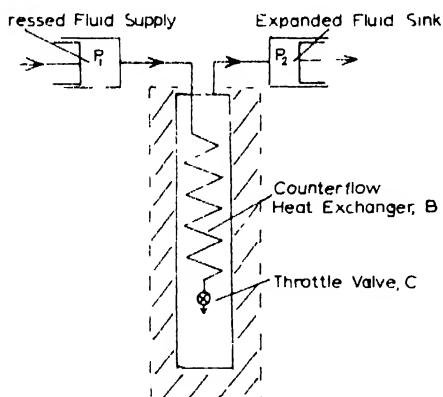


FIG. 2. Adiabatic throttling flow system.

constant high-pressure source of fluid at P_1 which flows at constant rate through an insulated heat exchanger B, then through a throttling device C and back through exchanger B, leaving the exchanger against some constant low pressure, P_2 , at the *same* temperature at which high-pressure fluid enters, gives

$$H_1 + Q = H_2$$

H_1 and H_2 are the enthalpies per unit mass of fluid entering and leaving the system, respectively, and Q is the amount of heat *absorbed* by a unit mass of fluid in passing through the system. The initial and final kinetic and gravitational potential energies are assumed equal. H_1 will be smaller than H_2 for most gases at absolute temperatures

less than 8 to 10 times their normal boiling point and for initial pressures of several hundred to a few thousand pounds per square inch. At higher temperatures, H_1 will be larger than H_2 so that Q is negative—heat is *liberated* within the exchanger-throttle valve system. If the thermal insulation of the exchanger-throttle valve system is *perfect*, then for $H_1 < H_2$, the exchanger-throttle valve system and the circulating gas itself will be continually cooled until some of the circulating gas accumulates within the system as liquid and a new energy balance $H_1 = fH_{\text{liquid}} + (1 - f)H_2$ is attained. The fraction, f , of the entering gas accumulates within the system as liquid product with enthalpy H_{liquid} , and the fraction $(1 - f)$ returns through the exchanger as its refrigerant.

Carl von Linde, in 1885, was the first to couple the relatively feeble (40°C maximum for air expanded from 4000 psi to one atmosphere room temperature) Joule-Thomson cooling, produced by throttling a high-pressure gas and a regenerative heat exchanger, to give the very simple system which is capable of liquefying any gas except helium, hydrogen and neon, starting from room temperature—albeit with poor efficiency. For helium, hydrogen and neon, throttling at room temperature produces heating ($H_2 > H_1$). Sir James Dewar used a bath of liquid oxygen to precool compressed hydrogen to $\sim 90\text{ K}$ where $H_2 < H_1$. The cooled hydrogen was then fed into a simple Linde liquefier, and Dewar first liquefied hydrogen in this way in 1898.

Kammerlingh Onnes used a bath of liquid hydrogen boiling under vacuum near its freezing point to precool compressed helium to $\sim 14\text{ K}$ (where $H_2 < H_1$ for helium). The helium was then fed into a simple Linde liquefier, and Onnes first liquefied helium in 1908.

For any real system, whose thermal insulation leaks q units of heat per unit mass of entering gas and whose exchanger permits gas to leave the system at $T_2' > T_1$, the energy balance becomes

$$H_1 + q = fH_{\text{liquid}} + (1 - f)[H_2 - C_p(T_2' - T_1)]$$

Poor insulation (large q) and an inefficient exchanger (large $T_2' - T_1$) can easily reduce f to zero.

As the temperature of the gas entering the exchanger of a Linde liquefier approaches its critical temperature, the thermodynamic efficiency of the system as a refrigerator rises sharply. A simple Linde liquefier (commonly but improperly called a Joule-Thomson liquefier) is combined with any of several types of efficient auxiliary preliminary refrigerators and forms the final stage of almost every large scale liquefier in common use. Linde, himself, quickly modified the simple system to what in essence is a pair of simple liquefiers operating in cascade. The first (precooling) unit operates between a common initial high pressure and some intermediate pressure for optimum efficiency, rather than between the high pressure and one atmosphere. He also added a conventional ammonia refrigerator for precooling to further enhance efficiency.

It is possible to produce a cold stream of gas with relatively good efficiency by allowing compressed gas to expand in a reciprocating expansion engine, or in an expansion turbine. If the expansion engine is preceded by an efficient regenerative heat exchanger, relatively modest ratios of inlet pressure to exhaust pressure at the expansion engine will produce gas near its boiling point. The thermodynamic efficiency of such expanders commonly approaches or exceeds 80 per cent. A fraction of this cold exhaust gas can be used to refrigerate the feed stream to a simple Linde liquefier. Georges Claude in 1905 combined a reciprocating expansion engine with a simple Linde liquefier as shown schematically in Fig. 3 to produce an air liquefier of improved efficiency which became known as the Claude cycle.

Peter Kapitza combined a precooling bath of liquid nitrogen in sequence with a reciprocating

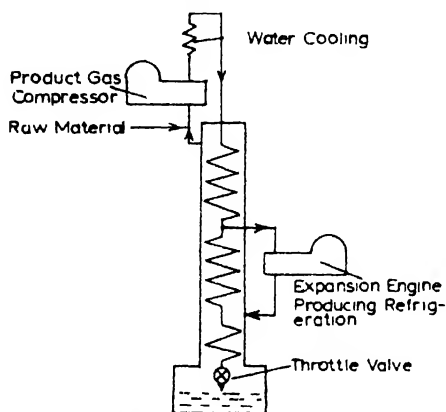


FIG. 3. Schematic arrangement Claude cycle.

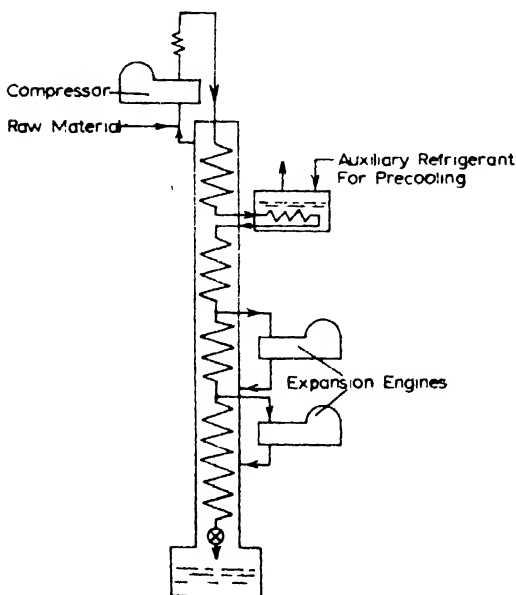


FIG. 4. Composite system for helium liquefaction.

expansion engine to precool compressed helium feed for a final Linde stage in 1934 and produced a Claude cycle liquefier which liquefied helium without the use of any auxiliary liquid hydrogen. Collins developed a similar machine which has been produced commercially in relatively large numbers. It uses a cascade of two expansion engines in a Claude cycle liquefier capable, when combined with liquid-nitrogen precooling, of producing ~8 liters of liquid helium per hour. The schematic arrangement is shown in Fig. 4. The same schematic arrangement of liquid precooling bath or baths followed by expansion engines all precooling the feed to a final Linde liquefier, could be used to describe in essence the large plants currently used for the liquefaction of hydrogen or helium.

DAVID N. LYON

References

- Vance, R. W., "Cryogenic Technology," Ch. 2, New York, John Wiley & Sons, 1964.
- Scott, R. B., "Cryogenic Engineering," Ch. 2, New York, D. Van Nostrand, 1959.
- Zemansky, M. W., "Heat and Thermodynamics," Third Edition, Ch. 14, New York, McGraw-Hill Book Co., 1951.
- Dodge, B. F., "Chemical Engineering Thermodynamics," Ch. 10, New York, McGraw-Hill Book Co., 1944.
- Collins, S. C., *Science*, **116**, 289 (1952)

Cross-references: CRYOGENICS, ENTROPY, GAS LAWS, REFRIGERATION, STATES OF MATTER, THERMODYNAMICS.

LIQUID STATE

Liquid is the term used for a state of matter characterized by that of a pure substance above the temperature of melting and below the vaporization temperature, at any pressure between the triple point pressure and the critical pressure (see Fig. 1). The liquid state resembles the crystalline in the relatively low dependence of density on P and T , and resembles the gas state in the inability to support shear stresses (see reference to glasses below). Structurally the molecules are relatively close together but they lack long-range crystalline order. The mutual solubility of different liquids is also intermediate between the complete mutual solubility of all gases, and the relatively rare appreciable mutual solubility of pure crystalline compounds. Two liquids of similar molecules are usually soluble in all proportions, but very low solubility is sufficiently common to permit the demonstration of as many as seven separate liquid phases in equilibrium at one temperature and pressure (mercury, gallium, phosphorus, perfluoro-kerosene, water, aniline, and heptane at 50°C, 1 atmosphere).

Stability Limits. With the exception of helium and certain apparent exceptions discussed below, Fig. 1 gives a universal phase diagram for all pure compounds. The triple point of one P and one T is the single point at which all three phases,

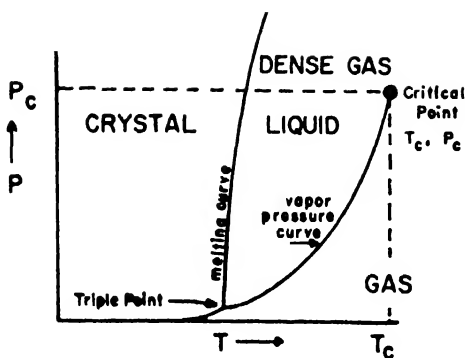


FIG. 1

crystal, liquid, and gas, are in equilibrium. The triple point pressure is normally below atmospheric. Those substances, i.e., CO_2 , $P_t = 3885$ mm, $T_t = 56.6^\circ\text{C}$ for which it lies above, sublime without melting at atmospheric pressure.

From the triple point, the melting curve defines the equilibrium between crystal and liquid, usually rising with small but positive dT/dP , and presumably always with positive dT/dP at sufficiently high P values. The line is believed to extend infinitely without a critical point (it has been followed to $T \approx 16T_c$ for He, and calculations indicate that hard spheres would show a gas-crystal phase change). The gas-liquid equilibrium line, the vapor pressure curve, has dT/dP always positive and greater than the melting curve. The vapor pressure curve always ends at a critical point, $P = P_c$, $T = T_c$, above which the liquid and gas phase are no longer distinguishable. Since the liquid can be continuously converted into the gas phase without discontinuous change of properties by any path in the P - T diagram passing above the critical point, there is no definite boundary between liquid and gas.

The term *liquid* is commonly reserved for $T < T_c$, and "dense gas" is used for $T > T_c$. However, certain properties, such as the ability to dissolve solids, change rather abruptly at the critical density. In many respects, the dense gas resembles the low-temperature liquid of the same density more closely than it does the dilute gas.

The slope, dT/dP , of all phase equilibrium lines obeys the thermodynamic Clapeyron equation:

$$dT/dP = \Delta V/\Delta S = T\Delta V/\Delta H \quad (1)$$

with ΔV , ΔS , and ΔH the differences, for the two phases, of volume, entropy, and heat content or enthalpy, respectively. The quantity ΔH is the heat absorbed in the phase change at constant P . Since always $S_{\text{cr}} < S_{\text{liq}} < S_{\text{gas}}$ and usually $V_{\text{cr}} < V_{\text{liq}} < V_{\text{gas}}$, one usually has $dT/dP > 0$; the relatively rare cases, including water, for which $V_{\text{liq}} < V_{\text{cr}}$, at low pressures leads to $dT/dP < 0$ for the melting curve near the triple point.

Figure 1 gives the P - T boundaries of the stable liquid phase. Clean liquids can readily be superheated or supercooled, and in vessels having walls

to which the liquid adheres, they can be made to support negative pressures of several tens of atmospheres. Thus the properties of the metastable liquid can be investigated outside the limits shown in the diagram.

Two apparent exceptions to the universality of the phase diagram of Fig. 1 deserve mention. First, many of the more complicated molecules decompose at temperatures below melting or boiling, and the diagram is unobservable. Secondly, some liquids, notably glycerine and SiO_2 and many multicomponent solutions, supercool so readily that crystallization is difficult to observe. In these cases, there is a continuous transition on cooling to a glass, which has the elastic properties of an isotropic solid. The structure of the glass is qualitatively that of the high-temperature liquid, lacking long-range order. Since glass and liquid are not sharply differentiated, the term *liquid* is sometimes used to include glasses, although common parlance reserves liquid for the state in which flow is relatively rapid.

Quantum Liquids. The one real exception to the phase diagram of Fig. 1 is that of helium, Fig. 2. Both isotopes, He^1 and He^3 , have no

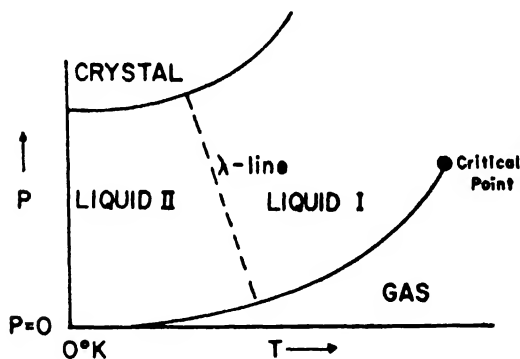


FIG. 2.

triple point, the liquid is stable to 0°K below about 20 atmosphere for He^1 and below about 30 atm for He^3 . The liquids have zero entropy at 0°K in both cases. This is also the only case in which isotopic mixtures form two liquid phases at equilibrium: the isotopic solution separating below 1 K. The isotope He^4 has itself two phases, He I above the dotted λ -line of the diagram, and He II with remarkable properties of superfluidity, second sound, etc., below the λ -line. The phase transition along the λ -line is second order; that is, whereas S and V are continuous, heat capacity and compressibility change discontinuously across the λ -line.

Although no completely satisfactory single theory of liquid helium has yet been formulated, one can say that most of the remarkable properties are qualitatively understood and are due to the predominance of quantum effects, including the difference in the statistics of the even and odd isotopes. Thus helium is the one example in nature of a quantum liquid, all other liquids

showing only minor deviations from classical behavior.

Structure. Considerable confusion in the description of liquid structure exists, due primarily to difficulties of precise formulation of verbal concepts. The geometric arrangement of any small number (say 10 to 12) of close lying molecules resembles the arrangement in the crystal, but the order rapidly disappears as larger groups are considered. Long-range order is lacking. The fact that numerical theories based on a lattice or cell structure have some success is evidence only that most properties depend on the configuration of near neighbors alone. Insofar as the arrangement of nearest neighbors is describable in terms of that of the crystal, the structure of the normal liquid is probably characterized best by a somewhat closer spacing than the crystal of the same molecules, the reduced density arising from a considerable number of vacancies in the lattice; the coordination number or number of nearest neighbors is lower than in the crystal. The exception is water, in which the low coordination number, 4, of the crystal, is increased by interstitial molecules in the liquid, leading to a higher density of the liquid.

Structural descriptions of this nature usually lack the possibility of precise formulation. It is, however, possible to define for any disordered array of molecules in three-dimensional space an arrangement of contiguous cells, each containing one and only one molecule, the faces of the cells being the loci of the midpoints of neighboring molecules. The statistics of the fraction of cells with n faces and of the distances of the faces from the molecules would give the fraction of molecules having a given number of nearest neighbors and the distance distribution of these in a precisely defined manner. Neither present experimental information nor present theories lend themselves to analysis in such terms.

The only clearly defined manner of describing liquid structure in use at present involves the concept of a set of probability density functions, ρ_n , for ascending numbers, n , of molecules. The function ρ_n depends on the vector coordinates $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$ of n molecules, and

$$\rho_n(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n) d\mathbf{r}_1, \dots, d\mathbf{r}_n$$

is defined as being the probability that in the liquid of definite P and T , there will be at any instant of time, one molecule at each position, \mathbf{r}_i , within the volume element, $d\mathbf{r}_i$. For a fluid, unlike a perfect single crystal, $\rho_i(\mathbf{r})$ is a constant independent of \mathbf{r} and equal to the number density: the number, ρ , of molecules per unit volume. The first significant member of the set is then the pair density function, $\rho_2(\mathbf{r}_1, \mathbf{r}_2)$, which depends only on the distance, $r = |\mathbf{r}_1 - \mathbf{r}_2|$, between the two molecules. At large distances $\rho_2(r \rightarrow \infty) = \rho^2$. This function can be found experimentally from the x-ray scattering intensities of the liquid (it is the three-dimensional Fourier Transform of the scattering intensity at angle θ vs $(4\pi/\lambda)/\sin(\theta/2)$). A typical plot is shown in Fig. 3. The area under

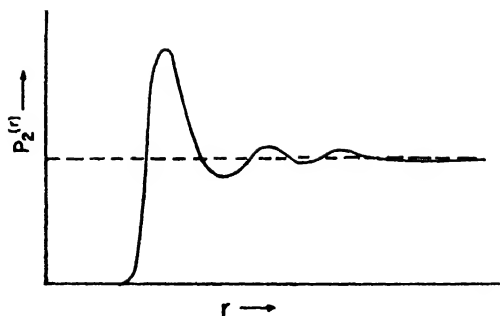


FIG. 3.

the ill-defined first peak integrated over $4\pi r^2 dr$ is the average number of nearest neighbors, and is of order 10 to 11 for normal liquids.

The quantity of dimensions of energy,

$$W_n(\mathbf{r}_1, \dots, \mathbf{r}_n) = kT \ln [\rho^{-n} \rho_n(\mathbf{r}_1, \dots, \mathbf{r}_n)]$$

can be shown to be the potential of average force of n molecules located at the positions $\mathbf{r}_1, \dots, \mathbf{r}_n$. That is, if there are n molecules at these positions there will be some average force, \bar{f}_i , along the x -coordinate of molecule i . This average is the sum of the direct force due to the other $n-1$ plus the average of a fluctuating force due to the others, whose average position is affected by that of the n specified ones. This average force is

$$f_{xi} = -(\partial/\partial x_i) W_n(\mathbf{r}_1, \dots, \mathbf{r}_n)$$

One frequently assumes that W_n is a sum of pair forces only,

$$W_n(\mathbf{r}_1, \dots, \mathbf{r}_n) = \sum_{n \geq i > j \geq 1} W_2(\mathbf{r}_{ij})$$

although this assumption is known to be only approximate. With this assumption, the pair average force potential, $W_2(\mathbf{r}_{ij})$, can be computed as the solution of an integral equation, and the solutions agree quite well with the experimental curves.

The knowledge of the complete set of functions ρ_n plus that of the intermolecular forces would permit the computation of all equilibrium properties of the liquid, and indeed if the intermolecular forces are the sum of pair forces, only a knowledge of ρ_2 at all P, T values is necessary. An adequate, although numerically difficult, theory of the transport properties also exists, using the equilibrium functions, ρ_n . At present, only qualitative success is obtained in the completely *a priori* use of the equations.

Associated Liquids. The description given above is adequate only for liquids composed of spherically symmetric molecules or molecules that are nearly so. These constitute the so-called normal liquids, which obey reasonably well the law of corresponding states, for which the entropy of vaporization at the boiling point has the Trouton's Rule value of approximately 21 cal/deg. For molecules containing large dipole moments,

or those forming mutual hydrogen bonds, the concept of the probability density functions must be extended to include angles or other internal degrees of freedom in the coordinates. Such inclusion is conceptually easy, but incredibly complicates the already difficult numerical evaluation of any equations. However, certain qualitative statements may be made.

Liquids composed of molecules with large dipole moments are frequently referred to as associated. Although in some instances relatively stable dimer or definite polymer units of relatively fixed orientation may exist, in many cases, notably water, it is extremely doubtful if an exact knowledge of the structure would reveal any distinguishable entities of associated molecules other than that of the whole liquid. In such cases, one would, however, expect that certain mutual angular orientations between neighboring molecules will be highly preferred, whereas in the dilute gas this will not be the case. The effect of this restriction on the internal coordinates will be to decrease the entropy of the liquid markedly compared to the gas. This effect is qualitatively the same as in association, and the properties of these liquids, particularly the high entropy of vaporization, will simulate those of a liquid composed of definite associated complexes.

JOSEPH E. MAYER

Cross-references: CRYSTALLOGRAPHY, DIPOLE MOMENTS, ENTROPY, SUPERFLUIDITY, SURFACE TENSION, THERMODYNAMICS, TRANSPORT THEORY, VAPOR PRESSURE AND EVAPORATION.

LUMINANCE

Light has been defined¹ as "the aspect of radiant energy of which a human observer is aware through the visual sensations which arise from the stimulation of the retina of the eye."

One can specify the *radiant intensity* of a point source in a given direction quite simply as so many ergs/second/steradian (cgs system), or as watts/steradian (mks system). Similarly one can specify the *radiance* of an extended source, i.e., the areal density of point sources of unit radiant intensity to which the extended source is equivalent, as so many ergs/second/steradian/square centimeter, or as watts/steradian/square meter.

The problem, however, is that the visual effect is not simply related to the amount of radiation expressed in physical terms. It is evident from observation of the spectrum of a source radiating equal amounts of energy per unit wavelength interval throughout the visible spectrum, that equal energy per unit wavelength interval does not produce visual sensations having equal brightness. The problem is further compounded by the fact that the visual effect is also dependent upon the size of the field viewed, the duration of stimulation to some extent, the part of the retina stimulated, the adaptation state, and the characteristics of the individual observer.

For given conditions of observation, however,

and a sufficiently large number of observers, one can determine the average *relative luminous efficiency* of a monochromatic radiation, i.e., the ratio of the luminous efficiency of that radiation to the maximum luminous efficiency, by a variety of indirect methods. The work of Gibson and Tyndall² is particularly relevant here as being the basis for the standards adopted by the International Commission on Illumination at Geneva in 1924.

Luminous flux can now be defined as "*that quantity characteristic of radiant flux which expresses its capacity to produce visual sensation, evaluated according to the values of relative luminous efficiency adopted by the International Commission.*"¹ The luminous equivalent of radiant intensity and radiance now become *luminous intensity* and *luminance*. The former is defined, in any direction, as *the ratio of the luminous flux emitted by a source or by an element of a source, in an infinitesimal cone containing this direction, to the solid angle of this cone.*³ The latter is defined, at a point of a surface and in any direction, as *the ratio of the luminous intensity in that direction of an infinitesimal element of the surface containing the point under consideration, to the orthogonally projected area of this element on a plane perpendicular to that direction.*⁴

The unit of luminous intensity is the *candela*, which is of such a value that the luminous intensity of a full radiator at the freezing point of platinum is 60 units of luminous intensity per square centimeter. The unit of luminous flux is the *lumen*, the flux emitted in a solid angle of one steradian by a uniform point source having an intensity of one candela.

Since luminance is expressed in terms of luminous intensity per unit projected area, the unit of luminance is the candela per unit area. The unit recognized internationally is the *nit*, the candela per square meter. The other unit in the metric system is the *stilb*, or candela per square centimeter. In the British system, the units used are the candela per square inch and the candela per square foot.

Luminance can also be assessed in terms of reflected or emitted luminous flux per unit area, and this is the rationale for the system of units based on the lumen per unit area. The primary unit in the metric system is the *lambert*, the unit of a perfectly diffusing surface, emitting or reflecting light at the rate of one lumen per square centimeter. Commonly used derivatives of the lambert are the *millilambert* (10^{-3} lambert), and the *microlambert* (10^{-6} lambert). Sometimes encountered is the *apostilb* (10^{-4} lambert), which is the luminance of an ideal diffuser emitting or reflecting one lumen per square meter. The *foot-lambert* is the luminance of an ideal diffuser emitting or reflecting one lumen per square foot.

Since an ideal diffuser with a luminance of one candela per unit area in all directions, emits π (3.1416) lumens per unit area, one candela per square centimeter will be equal to π lamberts; one candela per square foot will be equal to π foot-lamberts; and one nit will be equal to π apostilbs.

This multiplicity of units is unfortunate, but should not be confusing if the above relationships are remembered. The obvious solution is to use the most convenient system and convert when necessary. Table I should prove helpful.

P. J. FOLEY

TABLE I. CONVERSION FACTORS FOR UNITS OF LUMINANCE*

	cd/m^2	$cd/in.^2$	cd/ft^2	L	ml	<i>apostilb</i>	<i>ft-L</i>
cd/m^2 (nit)	1	1×10^{-4}	6.452×10^{-4}	3.142×10^{-4}	3.142×10^{-1}	3.142	2.919×10^{-1}
cd/cm^2 (srilb)	1×10^4	1	9.290×10^2	3.142×10^3	3.142×10^3	3.142×10^4	2.919×10^3
$cd/in.^2$	1.550×10^3	1	1.44×10^2	4.869×10^{-1}	4.869×10^2	4.869×10^3	4.524×10^2
cd/ft^2	1.076×10	1.076×10^{-3}	1	3.382×10^{-3}	3.382	3.382×10	3.142
lambert	3.183×10^3	3.183×10^{-1}	2.957×10^2	1	1×10^3	1×10^4	9.290×10^2
millilambert	3.183	3.183×10^{-4}	2.957×10^{-1}	1×10^{-3}	1	1×10	9.290×10^{-1}
apostilb	3.183×10^{-1}	2.054×10^{-4}	2.957×10^{-2}	1×10^{-1}	1×10^{-1}	1	9.290×10^{-2}
foot-lambert	3.426	3.426×10^{-4}	3.183×10^{-1}	1.076×10^{-3}	1.076	1.076×10	1

* Value in units in left column multiplied by the conversion factor equals value in units in upper row.

References

1. Committee on Colorimetry, Optical Society of America. *J. Opt. Soc. Am.* **34**, No. 5 (1944).
 2. Gibson, K. S., and Tyndall, E. P. T., "Visibility of Radiant Energy," *Natl. Bur. St., Sci. Papers*, **19** (1923).
 3. I.C.I. Definitions, Committee 1b, *J. Opt. Soc. Am.*, **41**, No. 10 (1951).
- Recommended, in addition to the above references:
 Wright, W. D., "Photometry and the Eye," London, Hatton Press Ltd., 1949.
 Walsh, J. W. T., "Photometry," London, Constable and Co. Ltd., 1953.
 Committee on Colorimetry, Optical Society of America, "The Science of Color," New York, Thomas Y. Crowell Co., 1954.

Cross-references: LIGHT; OPTICS, GEOMETRICAL; OPTICS, PHYSICAL; VISION AND THE EYE.

LUMINESCENCE

Introduction. Luminescence is the phenomenon of light emission in excess of thermal radiation. Excitation of the luminescent substance is prerequisite to the luminescent emission. Photoluminescence depends upon excitation by photons; cathodoluminescence, by cathode rays; electroluminescence, by an applied voltage; chemiluminescence, by utilization of the energy of a chemical reaction. Luminescent emission involves optical transitions between electronic states characteristic of the radiating substance. The phenomenon is essentially the emission spectroscopy of gases, liquids, and solids. The same basic processes may yield infrared or ultraviolet radiation in substances with suitable electronic energy states; therefore, such emission in excess of thermal radiation is also described as luminescence.

Luminescence can be distinguished from the Raman effect, Compton and Raleigh scattering and Cherenkov emission on the basis of the time delay between excitation and luminescent emission being long compared to the period of the radiation, λ/c , where λ is the wavelength and c is the velocity of light. The radiative lifetimes of the excited states vary from 10^{-10} to 10^{-1} second depending on the identity of the luminescent substances whereas λ/c is approximately 10^{-14} second for visible radiation. At ordinary densities of excitation, the spontaneous transition probability predominates so that the luminescent radiation is incoherent; under conditions of high densities of excitation in suitable luminescent substances, the induced transition probability may predominate, the emitted radiation is coherent, and laser action is attained.

The initial persistence of luminescent emission following the removal of excitation depends on the lifetime of the excited state. This emission decays exponentially and is often called fluorescence. In many substances, there is an additional component to the afterglow which decays more slowly and with more complex kinetics. This is

called phosphorescence. For many inorganic crystals, the emission spectra for fluorescence and phosphorescence are the same; the difference in afterglow arises from electron traps from which thermal activation is prerequisite to emission. For organic molecules, the emission spectra for fluorescence and phosphorescence are often different: the former occurs from an excited singlet; the latter, from a triplet state.

Luminescence of Gases. The simplest luminescent substances are monoatomic gases. The electronic states are characteristic of the isolated atoms; therefore, the excitation and emission spectra depend only on the differences in energy of the stationary electronic states of the many-electron atom, and the spectral lines are broadened only by the lifetimes of the excited states or at higher pressures, by collisions. The transitions are to a good approximation one-electron transitions. Resonance fluorescence is photoluminescence in which the exciting radiation is the exact frequency or wavelength for the transition from the ground to the excited state and emission occurs with the same frequency. Resonance fluorescence is shown diagrammatically in Fig. 1 for low pressure alkali

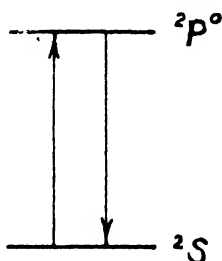


FIG. 1. Resonance fluorescence of atoms, e.g., Na.

metal vapor. The well-known 2537Å line of mercury vapor is another example of resonance fluorescence. This emission can also be excited by electrons accelerated by 5 or more volts. The simplest case of sensitized fluorescence (photoluminescence in which absorption of the exciting radiation is by one substance and the excitation is transferred to another which emits radiation) occurs with mixtures of monoatomic gases. For example, the characteristic fluorescence of thallium is observed when mixtures of Tl and Hg vapors are illuminated with the 2537Å radiation of Hg.

For diatomic and polyatomic gases, the energies of the electronic states are dependent on the interatomic distances of the molecule. This dependence is shown in Fig. 2 for the ground and excited states. For a diatomic molecule the coordinate R is the distance between the two atoms. For each electronic state, there is a series of vibrational levels which are also shown in Fig. 2. Optical transitions occur between individual vibrational levels of one electronic state and individual vibrational levels of another electronic state. These transitions occur in accordance with the Franck-Condon principle, i.e., with fixed

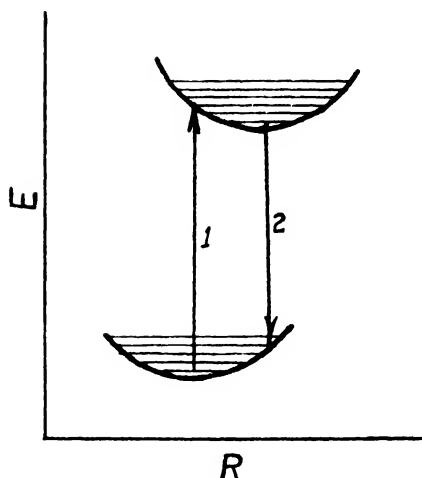


FIG. 2. Configuration coordinate model.

nuclear coordinates, vertically as shown in transitions 1 and 2 of Fig. 2. In most cases the emission will involve a smaller transition energy and occur at longer wavelength than the photoexcitation. This is referred to as Stokes' emission. In some cases, for example at high temperatures when the higher vibrational levels of the ground electronic state are populated thermally, anti-Stokes' emission is also observed. Additional structure in the photoexcitation and emission spectra arises from rotational states of the molecule. Iodine is a typical diatomic luminescent molecule excited by green light with visible emission at slightly longer wavelengths. Benzene and aniline are typical polyatomic molecules which are luminescent as vapors.

Luminescence of Organic Materials. The electronic states of most organic luminescent materials in the liquid or solid phase, either as pure materials or as solutes in dilute concentration in inert solvents, are to a good approximation describable in terms of the electronic states of the free molecule in the gaseous phase. In other words, the intermolecular forces are weak compared to the intramolecular forces. The photoexcitation and luminescent emission spectra of these substances in condensed phases are similar to the spectra of the vapors. The intermolecular forces are, however, great enough to bring about broadening of absorption and emission lines and in some cases to bring about electronic energy transfer between molecules before intramolecular, vibrational relaxation with the accompanying Stokes' shift can occur. On the other hand, in a viscous or rigid medium collisional, non-radiative de-excitation is reduced.

Many of the organic luminescent materials are aromatic molecules related to dyes. The sodium salt of fluorescein in dilute aqueous solution is well known as an efficient fluorescent material. Other organic substances luminesce efficiently when dissolved in organic solvents. Terphenyl in xylene is a liquid β - and γ -ray scintillator with emission in

the near ultraviolet. Some organic molecules luminesce most efficiently in a rigid medium. A solid solution of 1 per cent anthracene in naphthalene is a scintillator with blue emission. For these solutions, energy is absorbed by the solvent molecules and transferred to the solute where the luminescent emission occurs. Crystals of some pure organic substances luminesce, particularly at low temperatures.

The fluorescent emission and the long-wavelength absorption of organic materials are often simply related as mirror images of each other. This is shown in Fig. 3 for rhodamine in ethanol

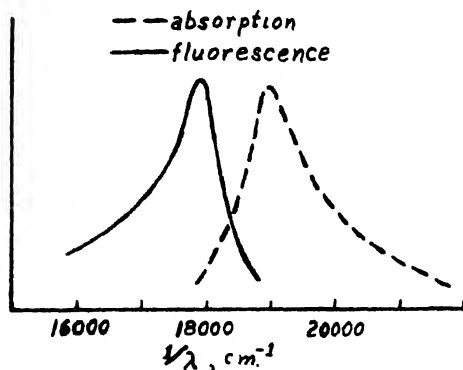


FIG. 3. Mirror symmetry of absorption and emission.

and can be explained on the basis of the configuration coordinate model in Fig. 2. with the force constants for the two electronic states approximately equal. For organic molecules, the configuration coordinate R is interpreted as representing, schematically, all the intramolecular nuclear coordinates. In addition to the fluorescent emission, many organic substances exhibit phosphorescent emission. This arises from non-radiative relaxation from the excited singlet to the triplet state followed by radiative decay from the triplet to the ground singlet state, as illustrated in Fig. 4. Because of the spin selection rule governing radiative transitions, the triplet has a long lifetime and the oscillator strength for direct excitation to the triplet is negligible. In suitable systems,

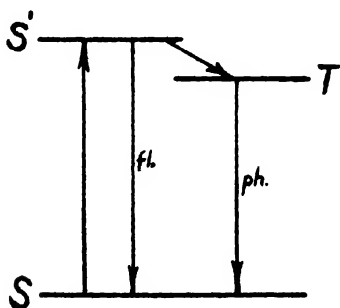


FIG. 4. Fluorescence and phosphorescence of organic molecules.

polarized excitation and emission arising from the anisotropy of the organic molecules can be observed.

Chelates involving organic molecules as ligands bound to a metal atom or ion are a class of substances with members which luminesce. The fluorescence and phosphorescence of chlorophyll both *in vivo* and *in vitro* have been investigated for many years. On the other hand, rare earth chelates with organic ligands have been intensively investigated quite recently as lasers. The photoexcitation occurs in the broad absorption bands of the ligand; the energy is transferred to the localized 4f shell of the rare earth by a mechanism related to the transfer process occurring in the scintillators; luminescent line emission characteristic of the rare earth occurs, as coherent radiation with high excitation intensities. More recently, a fluorinated Eu-acetate dissolved in acetonitrile has been announced as a liquid laser operating near room temperature.

Luminescence of Inorganic Crystals. Inorganic crystals which luminesce are often called phosphors. Their luminescence in most cases originates from impurities or imperfections. Exceptions include the luminescence of alkaline earth tungstates which is characteristic of the WO_4 group perturbed by the crystal field, the luminescence of some rare earth salts, and radiation recombination of conduction electrons with valence band holes in semi-conductors. The impurities and imperfections responsible for luminescence in inorganic crystals are of diverse atomic and molecular types whose characteristics depend on the structure of the defect and on the electronic structure of the pure crystal. In some cases, the electronic states involved in the luminescence can be described in terms of energy levels of the impurity ion perturbed by the crystal field; in other cases, in terms of the crystal band structure perturbed by the impurity. The existence of conduction bands in inorganic crystals, particularly in semiconducting crystals, introduces additional mechanisms for the excitation of luminescence and for phosphorescence. For example, suitable impurities can be excited by alternate capture of injected conduction electrons and valence band holes, thus providing one mechanism for electroluminescence; on the other hand, an excited luminescent impurity may lose an electron to another defect via the conduction band, and the thermal activation necessary for return to the luminescent impurity is responsible for phosphorescence.

The alkali halides are simple ionic crystals which become luminescent when doped with suitable impurities. Thallium substituted in dilute concentration at cation sites in potassium chloride has the absorption and emission shown in Fig. 5. The absorption bands involve the $^1S \rightarrow ^3P^0$, $^1P^0$ transitions of the free ion perturbed by crystal interactions; the principal emission band, $^3P^0 \rightarrow ^1S$, is similarly perturbed. The spectra can be understood qualitatively with the aid of Fig. 2, modified with a second excited state and with the configuration coordinate interpreted as symmetric

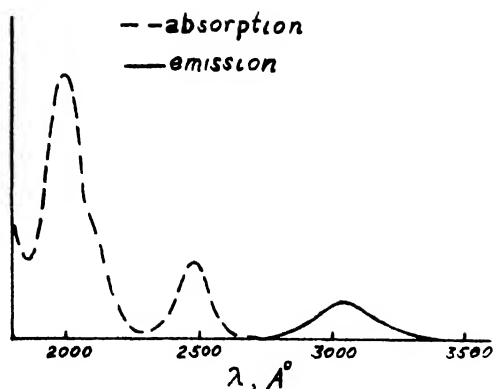


FIG. 5. Spectra of KCl:Tl.

displacement of the six nearest-neighbour Cl from the Tl^{+} . It is this interaction which is most dependent on the electronic state of the Tl^{+} and is, therefore, largely responsible for the band widths and Stokes' shift.

Many inorganic crystals become luminescent when certain transition metal ions are dissolved in them. The luminescence involves intercombination transitions within the $3d$ shell; therefore, crystal field theory can be used to interpret the absorption and emission spectra. Divalent manganese is a common activator ion. Zn_2SiO_4 , ZnS and $3Ca_3(PO_4)_2 \cdot CaF_2$ are important phosphors activated with Mn^{+2} . The last, activated also with Sb^{+3} , is the principal fluorescent lamp phosphor. The excitation at 2537\AA from the Hg discharge occurs at the Sb^{+3} , whose energy level structure is similar to that of Tl^{+} ; part of the energy is radiated in a blue band due to Sb^{+3} and part is transferred to the Mn^{+2} which is responsible for an orange emission band. The ruby laser involves the luminescence of Cr^{+3} in Al_2O_3 . Excitation occurs in a broad absorption band, the system relaxes to another excited state from which emission occurs in a narrow band.

Rare earth ions, particularly trivalent, in solid solution in inorganic crystals and in glasses exhibit the emission characteristic of transitions in the $4f$ shell. For examples, samarium, europium, and terbium give visible emission; neodymium, infrared; and gadolinium, ultraviolet. Because of their narrow emission bands, the rare earth activated phosphors are of interest as lasers and photon counters. Crystal field theory can be used to explain the optical absorption and luminescent emission of the $4f$ transitions of rare earth ions in crystals and glasses.

The zinc sulfide phosphors, which are widely used as cathodoluminescent phosphors and well-known for their electroluminescence, are now recognized as large band gap, compound semiconductors. Two impurities or imperfections are essential to the luminescence of many of these phosphors; an activator which determines the emission spectrum and a coactivator which is essential for the emission but in most cases has no effect on the spectrum. Activator atoms such as

Cu, Ag and Au substitute at Zn sites and perturb a series of electronic states upward from the valence band edge. In a neutral crystal containing only these activator impurities, the highest state is empty, i.e., it contains a positive hole and can accept an electron from the valence band; therefore, in semiconductor notation, the activator is an acceptor. In a similar way, coactivators such as Ga or In at Zn sites or Cl at S sites are donors. The simultaneous introduction of both types of impurities results in electron transfer from donor to acceptor lowering the energy of the crystal and leaving both impurities charged. The coulomb attraction of the donor and acceptor leads to a departure from a random distribution over lattice sites and to pairing. The electronic states and some of the transitions of acceptors, donors and donor-acceptor pairs are shown in Fig. 6. The spectrum of $ZnS:Cu, Ga$ is shown in Fig. 7. The longer-wavelength emission band involves the transition from the lowest donor state to highest acceptor state (transition 3) in approximately fifth nearest-neighbor pairs; the shorter-wavelength emission corresponds more nearly to transition 1 of Fig. 6. Luminescent emission from donor-acceptor pairs has been more clearly seen with gallium phosphide crystals. In addition to luminescence due to donors, acceptors and their pairs, emission bands due to transition metals are well known for zinc sulfide as noted earlier. In zinc sulfide crystals, the

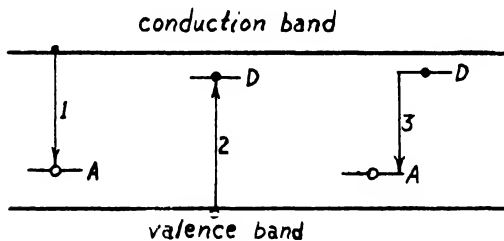
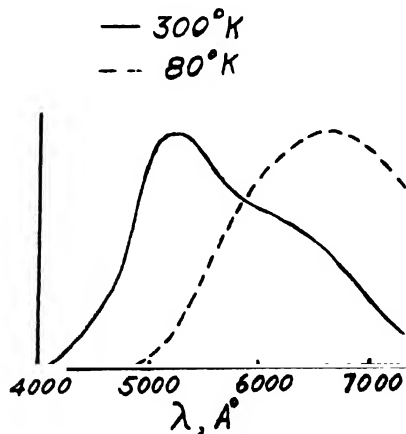


FIG. 6. Band model for acceptor, donor, and pair transitions.

FIG. 7. Spectra of $ZnS: Cu, Ga$.

donors which are unassociated with acceptors serve as electron traps and are responsible for long-persistent, temperature-dependent phosphorescence.

FERD WILLIAMS

References

Pringsheim, P., "Fluorescence and Phosphorescence," New York, Interscience Publishers, 1948.

Förster, T., "Fluoreszenz Organischer Verbindungen," Göttingen, Vandenhoeck and Ruprecht, 1951.

Kallmann, H., and Spruch, G., Eds., "Luminescence of Organic and Inorganic Materials," New York, John Wiley & Sons, 1962.

Curie, D., "Luminescence in Crystals," J. Wiley, New York, John Wiley & Sons, 1963 (translated by G. F. J. Garlick).

Cross-references: COLOR CENTERS; CRYSTALLOGRAPHY; ENERGY LEVELS; LASERS; PHOTOCONDUCTIVITY; RADIATION, THERMAL; RESONANCE; SEMICONDUCTORS; SOLID STATE PHYSICS; SOLID STATE THEORY; SPECTROSCOPY.

M

MAGNETIC FIELD

Basic Equations. To separate electromagnetic theory from the theory of the solid state, Maxwell's equations can be written in terms of the magnetic induction or flux-density \mathbf{B} and the total current density \mathbf{J} :

$$\nabla \cdot \mathbf{B} = 0 \quad (1)$$

$$\nabla \times \mathbf{B} = 4\pi\mathbf{J} \quad (2)$$

where, for purposes of the present article, the displacement current $(1/c^2)\delta\mathbf{E}/\delta t$ is neglected relative to $4\pi\mathbf{J}$ (see ELECTROMAGNETIC THEORY and MAGNETISM).

The solution of Eqs. (1) and (2) is

$$\mathbf{B}(\mathbf{r}) = \int d^3r_1 \frac{\mathbf{J}(\mathbf{r}_1) \times (\mathbf{r} - \mathbf{r}_1)}{|\mathbf{r} - \mathbf{r}_1|^3} \quad (3)$$

where the integral includes all current-carriers, and \mathbf{B} vanishes at infinity. In terms of the vector potential \mathbf{A} ,

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (4)$$

and the gauge $\nabla \cdot \mathbf{A} = 0$, one has

$$\nabla^2 \mathbf{A} = -4\pi\mathbf{J} \quad (5)$$

and

$$\mathbf{A}(\mathbf{r}) = \int d^3r_1 \frac{\mathbf{J}(\mathbf{r}_1)}{|\mathbf{r} - \mathbf{r}_1|} \quad (6)$$

for \mathbf{A} vanishing at infinity.

Magnetic lines of force are defined by $d\mathbf{r} \propto \mathbf{B}$ and are endless, by virtue of Eq. (1). The magnetic flux Φ through a surface S , bounded by a closed curve ℓ , is given by

$$\Phi = \int dS \mathbf{B} \cdot \mathbf{n} = \oint d\ell \cdot \mathbf{A} \quad (7)$$

where \mathbf{n} is the normal to S . The flux tube defined by the field lines passing through ℓ contains constant flux, independent of S .

At a large distance from a localized current distribution at $\mathbf{r} = 0$, Eq. (6) gives the dipole potential

$$\mathbf{A} = \frac{\mathbf{m} \times \mathbf{r}}{r^3} \quad (8)$$

where

$$\mathbf{m} = \frac{1}{2} \int d^3r_1 \mathbf{r}_1 \times \mathbf{J}(\mathbf{r}_1) \quad (9)$$

Equations in the Presence of Magnetic Materials. A macroscopic current density \mathbf{J}_M can be defined by local averaging of the microscopic current density \mathbf{J}_m within magnetic materials

$$\mathbf{J}_M = \langle \mathbf{J}_m \rangle_{av} \quad (10)$$

More conveniently, a magnetization vector \mathbf{M} can be introduced, where

$$\mathbf{J}_M = \nabla \times \mathbf{M} \quad (11)$$

In experiments, only \mathbf{J}_M can be measured directly (via measurements on \mathbf{B}), and \mathbf{M} is then uniquely derivable from Eq. (11) only with the added condition that it is to be a local state variable of the magnetic material (i.e., it is constant in uniform samples and constant fields). This condition follows automatically from the interpretation of \mathbf{M} as a magnetic-moment density per unit volume:

$$\mathbf{M} = N\mathbf{m}_0 \quad (12)$$

The theoretical molecular magnetic moment \mathbf{m}_0 (with number density N) is derived from \mathbf{J}_m by evaluation of Eq. (9) over the molecular volume.

For macroscopic purposes, the total current density \mathbf{J} of the preceding section is now specified by

$$\mathbf{J} = \mathbf{J}_c + \mathbf{J}_M \quad (13)$$

The component \mathbf{J}_c flows in conductors of resistivity η in accordance with Ohm's Law:

$$\eta \mathbf{J}_c = \mathbf{E} \quad (14)$$

The component \mathbf{J}_M is derived from \mathbf{M} .

In the analysis of configurations involving magnetic materials, the magnetic field \mathbf{H} is a convenient vector

$$\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M} \quad (15)$$

Then Eqs. (1) and (2) take the form

$$\nabla \cdot \mathbf{H} = -4\pi \nabla \cdot \mathbf{M} \quad (16)$$

$$\nabla \times \mathbf{H} = 4\pi \mathbf{J}_c \quad (17)$$

At the interface between two magnetic materials, Eqs. (1) and (17) imply continuity of the normal component of \mathbf{B} and of the tangential component of \mathbf{H} .

Across a sheet-current of density I_c/ℓ per unit length, the tangential component transverse to

\mathbf{J}_c of both \mathbf{H} and \mathbf{B} undergoes an increment $4\pi I_c/l$. The other components of \mathbf{H} and \mathbf{B} are unaffected.

For weakly magnetic materials, Eq. (15) can generally be written in terms of a scalar magnetic permeability μ

$$\mu \mathbf{H} = \mathbf{B} \quad (18)$$

For ferromagnetic materials, one can still write

$$\mu \mathbf{H} = \mathbf{B} - 4\pi \mathbf{M}_0 \quad (19)$$

where \mathbf{M}_0 is a permanent magnetization, but μ now depends on the time history as well as the magnitude of \mathbf{H} .

When \mathbf{J}_c is zero everywhere, one can define a scalar potential Ω , such that

$$\mathbf{H} = -\nabla \Omega \quad (20)$$

$$\nabla^2 \Omega = 4\pi \nabla \cdot \mathbf{M} \quad (21)$$

If the boundary condition on Ω is simply that it vanish at infinity, the solution is

$$\Omega(\mathbf{r}) = -\int d^3 r_1 \frac{\nabla_1 \cdot \mathbf{M}(\mathbf{r}_1)}{|\mathbf{r} - \mathbf{r}_1|} \quad (22)$$

In the presence of current-carrying conductors Eqs. (20) and (21) still hold in the region where $\mathbf{J}_c = 0$, but Eq. (17) now implies a multivalued potential

$$\oint d\mathbf{l} \cdot \mathbf{H} = \oint d\Omega = 4\pi I_c \quad (23)$$

where the integral is taken around a loop enclosing the total conductor current I_c . To keep the potential single-valued, so that the solution of Eq. (22) remains valid, one may adopt the "magnetic-shell" approach: \mathbf{J}_c is replaced with an equivalent \mathbf{M} , in analogy with Eq. (11).

Magnetic Force and Energy. From Maxwell's stress tensor, we find the volume force

$$\mathbf{f} = -\nabla \left(\frac{B^2}{8\pi} \right) + \frac{1}{4\pi} (\mathbf{B} \cdot \nabla) \mathbf{B} \quad (24)$$

$$= \mathbf{J} \times \mathbf{B}$$

which agrees with the summation of the Lorentz forces on the moving charges composing \mathbf{J} . The "magnetic pressure" against a current sheet bounding a region of finite \mathbf{B} (as in the Meissner effect or ordinary skin effect) is thus $B^2/8\pi$, evaluated at the surface. The force and torque on a body localized in a nearly uniform field are

$$\mathbf{F} = (\mathbf{m} \cdot \nabla) \mathbf{B} \quad (26)$$

$$\mathbf{N} = \mathbf{m} \times \mathbf{B} \quad (27)$$

From the microscopic point of view underlying Eq. (2), the magnetic energy density is

$$w = \frac{B^2}{8\pi} \quad (28)$$

In the presence of magnetic materials, one is more interested in the electrical input energy

required to go from \mathbf{B}_0 to \mathbf{B} , and this is given by

$$\Delta w = \frac{1}{4\pi} \int_{\mathbf{B}_0}^{\mathbf{B}} \mathbf{H} \cdot d\mathbf{B} \quad (29)$$

For $\mathbf{H} = \mu \mathbf{B}$, with constant μ , this becomes

$$\Delta w = \frac{1}{8\pi\mu} (B^2 - B_0^2) \quad (30)$$

The derivation of Eqs. (28) and (29) depends on the complete set of Maxwell's equations.

HAROLD P. FURTH

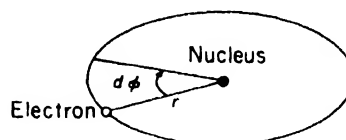
References

- Stratton, J. A., "Electromagnetic Theory," New York, McGraw-Hill Book Co., 1941.
Jackson, J. D., "Classical Electrodynamics," New York, John Wiley & Sons, 1962.

MAGNETIC RESONANCE

Electrons revolve about the nucleus of an atom and spin around their axes; in addition, the nucleus has a spin of its own. All of these moving charges have associated magnetic fields (magnetic moments), and *magnetic resonance* is concerned with the interactions of some of these fields with each other and with at least two external magnetic fields applied to the atom.

To facilitate understanding, consider an electron circulating about the nucleus. The electron has angular motion and it is a charged particle held in orbit by the oppositely charged



nucleus. Assume KEPLER'S LAW of areas applies (radius vector of the particle sweeps out equal areas in equal times) and that angular momentum, P_ϕ , is conserved and quantized ($P_\phi = L\hbar = mr^2\dot{\phi}$); $L \equiv$ orbital quantum number, and $\hbar \equiv$ Planck's constant/ 2π . The area swept out in one period, t , is:

$$A = \int_0^t \frac{1}{2} (r^2 \dot{\phi}) dt = L\hbar t / 2m.$$

From Amperes Law, one can find the equivalent magnetic dipole moment, μ , produced by a closed current loop, I , to be: $\mu = I \cdot A$. Since current is e/t it follows that

$$\mu = L(e\hbar/2m) \quad (1)$$

and $e\hbar/2m$ is defined as a Bohr magneton, μ_B .

Next, consider the ELECTRON SPIN about its own axis. A classical derivation of the spin moment similar to the above produces an electron spin

moment of two Bohr magnetons. Dirac's relativistic quantum theory of the electron and many experiments give the correct value of one Bohr magneton; this means the spin quantum number of the electron is $1/2$.

The total angular momentum, P_J , is found by adding vectorially the orbital and spin angular momentum of the electron, $P_J = P_L + P_S = \hbar(L + S)$ and the total magnetic moment becomes: $\mu = \mu_B(L + 2S)$. P_L and P_S can be thought of as precessing about P_J and only their components in the direction of P_J contribute to the average magnetic moment. Vectorially adding the moments gives:

$$\begin{aligned}\mu_J &= \mu_B [L \cos(LJ) + 2S \cos(SJ)] \\ &= \mu_B J \left(1 + \frac{S^2 + J^2 - L^2}{2J^2} \right)\end{aligned}$$

or using a more rigorous wave mechanics approach S^2 is replaced by $S(S + 1)$, etc., to give:

$$\mu_J = \mu_B gJ \quad (2)$$

where

$$g \equiv 1 + \frac{S(S + 1) + J(J + 1) - L(L + 1)}{2J(J + 1)}$$

The quantity g is called the Landé g -factor and for an atom in the ground state $L = 0$, $S = J$, and g becomes equal to 2. We could add the nuclear spin and its magnetic moment to this and the vector problem would become complex indeed! This will be done later in an easier way.

For simplicity, we will now add the external magnetic field, H . Just as a spinning top will precess in the earth's gravitational field, so will the magnetic moment vector of the electron precess in the magnetic field, the torque in the electron case being produced by the interactions of the dipole and the external field. Equating the time rate of change of angular momentum to the torque on the dipole, one can derive the precessional frequency in complete analogy to the top problem.

This precessional frequency, called the Larmor frequency, can also be derived from an energy standpoint and will give us more insight. The potential energy of a magnetic dipole in a magnetic field is $W = -\mu H = \mu_B g J H$. If we confine ourselves to an atom in the ground state then $J = S$ and $S = \pm 1/2$, the spin being either parallel or antiparallel with the external field. The magnetic moment is defined as positive or negative according to the condition of parallelism or antiparallelism, respectively.

Thus the energy difference between the two possible electron spin states can be equated to $\hbar\omega_L$, where ω_L is the Larmor frequency of precession.

$$\hbar\omega_L = W(S = \frac{1}{2}) - W(S = -\frac{1}{2}) = g\mu_B H = \gamma H \quad (3)$$

Electromagnetic radiation at the Larmor frequency and with the correct polarization will be absorbed by dipoles in the lower state, making

transitions to the higher state. This is shown in Fig. 1(b). *Electron spin resonance* is the technique for measuring this splitting using radio-frequency technology.

Since the nucleus is known to carry a charge, its angular spin should and does produce a nuclear magnetic moment. A nuclear magneton, μ_n , is defined in the same manner as the Bohr magneton, except the mass of the electron is replaced by the mass of the proton. A nuclear g -factor ($g_n \equiv gI$) is also defined where I is the spin of the nucleus.

The proton's magnetic moment is 2.7935 nuclear magnetons while the neutron's moment is $-1.9135 \mu_n$. The positive or negative sign refers to the condition of whether the angular momentum vector has the same or opposite direction as the magnetic moment. A nucleus with a spin, I , will have $2I + 1$ possible orientations in a magnetic field and corresponding $2I + 1$ energy levels. These give rise to a hyperfine structure.

For simplicity, consider a hydrogen atom in a molecule. The nucleus is a proton with a spin of $1/2$ and, consequently, its magnetic moment is either parallel or antiparallel to the field. This produces energy levels as shown in Fig. 1(c).

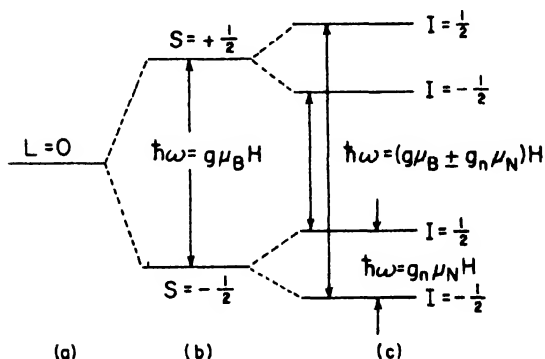


FIG. 1. Magnetic resonance energy levels. (a) orbital electron energy level; (b) electron spin moment in a magnetic field; (c) nuclear spin magnetic moments in a magnetic field.

Nuclear magnetic resonance is concerned with the use of radio-frequency techniques to study the transitions between nuclear spin states; in the case shown $\hbar\omega = g_n \mu_n H$ for the transition.

Consider a system of free magnetic dipoles in an external magnetic field. The magnetic dipoles would seek a position of minimum potential energy and consequently turn parallel with the field. In the world of nuclei and atoms we do not have such a closed system. The thermal agitation of the molecules and atoms coupled to the nuclear spin through magnetic and electric field interactions produces a disorienting effect which tends to fill all energy states equally.

The energy for each degree of freedom of a molecule is $(\frac{1}{2})kT$ and at room temperatures is roughly 10^{-14} ergs while the difference in energy between the two states of a proton (hydrogen nucleus) in a magnetic field of 10,000 oersteds is about 10^{-20} ergs ($W = g\mu_n H$).

For some solids and for liquids near absolute zero, more sophisticated distributions must be used, but for liquids and gasses at room temperature, a simple Boltzmann distribution will be adequate, $N \propto e^{-E/kT}$. Thus, the ratio of the populations of the two states used for an example would be $N(\frac{1}{2})/N(-\frac{1}{2}) \approx e^{(g\mu_n H/kT)}$ and for these protons at room temperature, one would find about seven more in the lower energy state out of every million.

As in electron resonance experiments, when an rf field is applied to the sample at right angles to the external field, an in-phase component of the rf field will cause the protons in the lower energy state to "flip" into the higher and the protons in the higher energy state into the lower.

In practice only, a certain percentage of each population is "flopped." However, a simple calculation will show that the population in both states will soon be practically the same; this condition is called saturation. The spin-lattice relaxation time, T_1 , is defined as the time during which all but $1/e$ of the excess spins will have reached the final equilibrium state of lowest energy.

Let A_0 be the magnitude of the energy absorbed to reach saturation; then A_0 is proportional to $g\mu_n H/kT_0$ where T_0 is the absolute temperature of the sample. If resonance absorption is permitted to take place before equilibrium of the excess occurs, then the energy absorbed is defined by a new T . This leads to the concept that resonance elevates the spin temperature, or heats up the spins; and after resonance stops, the spins cool to the temperature of the lattice. The lattice is defined as everything in the sample but the spin in question. T_1 for nuclear moments varies from hours to milliseconds, and for oxygen-free water at room temperature, it is about 3.6 seconds.

Another relaxation effect is that associated with the spin-spin interaction. Imagine an effective inhomogeneity in the external field of magnitude ΔH to exist over the region of a sample of the nuclear moments being studied. There will be variations in the external field at different nuclei, and we obtain in effect a spread in the Larmor frequencies throughout the sample, $|\Delta\omega| \approx |\gamma| |\Delta H|$. Spin-spin interactions are concerned with the effect of each precessing moment on the neighboring spins through their magnetic fields. Thus, the total field at each nucleus consists of the external field value and the resultant of the local fields produced by the components of the neighboring magnetic dipoles. T_2^* (transverse, spin-spin, or phase relaxation time) is defined as the time characteristic of the spread in Larmor frequencies. T_2 is the time during which all but $(1/e)$ of a system of spins whose phase has been established return to a state of random phase.

If the external field is quite homogeneous, the fields produced by neighboring dipoles are sufficient to actually separate the resonance absorption spectra of the molecule into patterns which identify the molecule. A very simple example of this is shown in Fig. 2 where the magnetic resonance spectra of ethyl alcohol is shown. In

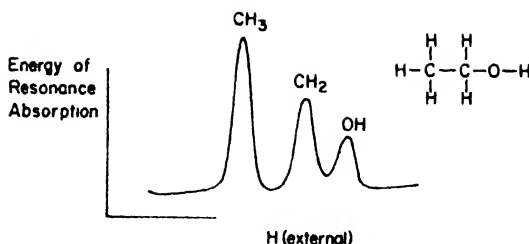


FIG. 2. Nuclear magnetic spectra of ethyl alcohol. Experimentally, the rf radiation on the atoms is usually held constant while the external magnetic field is varied. Absorption takes place when the Larmor frequency coincides with the radio frequency.

each of the three peaks, the proton of the hydrogen atoms is causing the resonance absorption.

A little thought and the use of symmetry arguments will show that the largest peak belongs to CH_3 , the next largest to CH_2 , and the smallest absorption to OH . Nuclear magnetic resonance (NMR) has become a very powerful tool for organic chemists. Magnetic resonance is also used to study free radicals, crystalline orientation, and many other things concerning the structure of matter.

JAMES T. SMITH

Cross-reference: ELECTRON SPIN, MAGNETISM.

MAGNETISM

Magnetization. Magnetic fields are produced both by macroscopic electric currents and by magnetized bodies. The first observed manifestations of magnetism were the forces between naturally occurring permanent magnets, and between these and the earth's field. North- and south-seeking poles could be identified. Poles were observed to be localized near the ends of long rods magnetized by contact with natural magnets or by a current-carrying coil. From the observed attraction and repulsion of unlike and like poles with an inverse square law came the concept of pole strength and the definition of the unit pole, that which acts on another in vacuum with a force of one dyne at a distance of one centimeter. The unit magnetic field, the oersted, could then be defined as that in which a unit pole experiences a force of one dyne. The magnetic moment of a long, uniformly magnetized rod of length l with a pole strength of m unit poles at each end is defined as ml , the largest couple that the sample can experience in a field of one oersted. The magnetization, M , is defined as the magnetic moment per unit volume, ml/al , where a is the cross-sectional area, and is thus also equal to the pole strength per unit area, m/a .

Magnetic Induction. The induction, or flux density, B , is numerically equal to the field H in free space and is described as one line of flux per square centimeter for a field of one oersted. Its direction is that of the force on a unit north pole.

If magnetic material is present, the flux density is equal to $H + 4\pi M$, since 4π lines of force emanate from the unit pole at each end of a specimen of unit magnetization. Magnetic poles are observed to occur in pairs. Lines of B are continuous, i.e., $\text{div } B = 0$. If a material becomes strongly magnetized in a small field, the lines of flux can be considered to crowd into the material, leaving their original locations and reducing the field there. This is how magnetic shielding is accomplished. Changes of B within a coil induce voltages which can be measured and form the basis of galvanometer and fluxmeter measurement methods. For a coil of N turns of cross-sectional area a , in which the flux is changing at dB/dt gauss/sec, E in volts is given by

$$E = 10^{-8} Na \frac{dB}{dt}$$

Forces on magnetic bodies in field gradients are proportional to M .

Types of Magnetic Behavior. In general a field H will produce a magnetization M in any material. If M is in the same direction as H , a sample will be attracted to regions of stronger field in a field gradient. It will be repelled if M is in the opposite sense. This experiment, as first performed by Faraday, is the basis for the broad classification of materials into paramagnetic, diamagnetic, and ferromagnetic. The susceptibility κ is defined as M/H . The force F_z on a small specimen of volume v in a field H_y and a field gradient dH_y/dx is

$$F_z = (\kappa_2 - \kappa_1)vH_y \frac{dH_y}{dx}$$

where κ_1 and κ_2 are the volume susceptibilities of the specimen and the surrounding medium, usually air.

For paramagnetic materials, κ is small and positive, usually between 1 and 1.001 at ordinary temperatures. These substances contain atoms or ions with at least one incomplete electron shell, giving them a non-zero atomic or ionic magnetic moment μ_a . Many salts of the iron-group and rare-earth metals are paramagnetic, as are the alkali metals, the platinum and palladium metals, carbon, oxygen, and various other elements. Antiferromagnetic substances also have small positive κ , as do ferromagnetics above their Curie temperatures. In the classical theory of paramagnetism, the orientations of the moments are considered to be initially thermally randomized in space. An applied field produces a net magnetic moment in its direction, as described by the classical Langevin function

$$\frac{M}{M_s} = \coth\left(\frac{\mu_a H}{kT}\right) - \frac{kT}{\mu_a H}$$

where k is the Boltzmann constant. M_s is the value of M attained for very large H/T . Under most conditions, only the initial portion of this curve is observed, with the corresponding constant κ . The conduction electrons at the top

of the Fermi distribution in a metal can also give rise to a temperature-independent Pauli paramagnetism. The quantum-mechanical analogue of the Langevin function is called the Brillouin function (see PARAMAGNETISM).

For diamagnetic materials, κ is small and negative. Diamagnetism is a universal phenomenon but is often masked by paramagnetic or ferromagnetic effects. Net diamagnetic behavior is observed in a number of salts and metals, and in the rare gases, in which there is no net moment. The effect can be regarded as the operation of Lenz's law on an atomic scale (see DIAMAGNETISM).

Ferromagnetic materials show a value of M which may be of the order of 10^3 in small fields. Thus κ can be very large. It is common to describe their properties in terms of the permeability $\mu = B/H$. Since M saturates in ordinary fields, κ and μ are not constant. M is not necessarily in the same direction as H , so κ and μ are, in general, tensors. Furthermore, ferromagnetics generally exhibit hysteresis in the dependence of M on H , and the details are very structure-sensitive. Still another distinction is the rather abrupt disappearance of ferromagnetism at a characteristic temperature, the Curie temperature, T_c .

Atomic Magnetic Moments. There are two possible sources for the moments of individual atoms. They are electron orbital motion and electron spin (see ATOMIC STRUCTURE). In ferromagnetic materials, most of the moment comes from spin rather than orbital motion, a fact that is revealed experimentally by gyromagnetic measurements (see FERROMAGNETISM) and by magnetic resonance experiments (see MAGNETIC RESONANCE). The unit of atomic moment is the Bohr magneton, μ_B , which is the moment associated with one electron spin, numerically equal to 0.9274×10^{-20} erg/oersted. The spin quantum number, S , is one-half the number of unpaired electrons. The moment per atom is $g\mu_B S$, where g is the gyromagnetic ratio, close to 2 for most materials. The moment in Bohr magnetons of an isolated atom or ion of the first transition series is equal to the number of unpaired d electrons, considering the first five electron spins to have one orientation and the next five the opposite (Hund's rule). The Ni^{++} ion, with eight d electrons, has the expected moment of $2\mu_B$ in ferrites, in which the ionic spacing is great enough so that the d levels are not disturbed (see FERRITES). In metallic nickel, however, the d levels overlap considerably, and the moment corresponds to only $0.6\mu_B$ per atom. Similarly the Bohr magneton numbers for metallic iron and cobalt are 2.2 and 1.7 respectively.

Ferromagnetism. Ferromagnetism can only occur in a material containing atoms with net moments. Also, quantum-mechanical electrostatic "exchange" forces must be present, holding neighboring atomic moments parallel below the Curie temperature. These are much greater than the Lorentz force due to the average magnetization and are, in fact, equivalent to an effective field on the order of 10^6 oersteds. Such an effective "molecular field" was postulated in 1907 by

Weiss in extending the Langevin theory of paramagnetism to include ferromagnetic behavior. The Langevin function predicts a temperature dependence of magnetization, for small M , of

$$M = \frac{CH}{T}$$

where C is a constant. The susceptibility is then C/T , which is Curie's law. Weiss pointed out that if the field H were augmented by an additional field NM proportional to the magnetization, the temperature dependence became

$$M = \frac{CH}{T - T_c}$$

where $T_c = NC$. This is the Curie-Weiss law, approximately obeyed by ferromagnetic substances above their Curie points. Below T_c , the presence of the molecular field produces an alignment of the atomic moments corresponding to the spontaneous magnetization M_s even when no external field is present. However, ferromagnetic materials can have any value of magnetization, including zero, which seems to contradict this result. Weiss therefore postulated the existence of domains separated by boundaries. In each domain the atomic moments are parallel, the domain magnetizations having different orientations. The net external magnetization is then the vector sum of the domain magnetizations and can be varied by a rearrangement of the domain structure, which may happen in very small applied fields. This prediction has been completely verified by experiment. The motion of domain boundaries as observed under the microscope has been directly correlated with external changes in magnetization. Domain boundaries in iron are on the order of 1000 Angstroms thick. Within a boundary, neighboring magnetic moments are not quite parallel. The change in orientation of the magnetization from one domain to another is distributed through the thickness of the boundary.

Within a domain, the magnetization will in general preferentially lie along some particular crystallographic direction. The energy difference between magnetization in the easiest and hardest direction may approach 10^7 ergs/cm³. This anisotropy is described in an appropriate trigonometric series with coefficients K_i . Usually only a few terms are necessary. Often a material is described by a single K ; this implies a uniaxial anisotropy energy of the form $K \sin^2\theta$. The K_i may pass through zero and change sign with changing composition or temperature. Although such details cannot in general be predicted, the magnetocrystalline anisotropy will have the same over-all symmetry as the crystal structure. Anisotropy is best investigated in single crystals, by analysis of magnetization curves in various directions or from the relationship between the measured torque and the direction of the applied field (see FERROMAGNETISM). Dimensional changes are also associated with the position of the magnetization vector relative to the lattice (see MAGNETOSTRICTION).

There are two mechanisms available for changing the externally measured magnetization of a ferromagnetic material: domain boundary motion, and domain magnetization rotation. Broadly speaking, in magnetically soft materials, boundary motion accounts for most of the changes in low applied fields, leaving the magnetization in each domain in the easy direction nearest the applied field. Then rotation against anisotropy produces the remaining change in higher fields. In very low fields, boundary motion is practically reversible, but when boundaries move considerable distances, they experience a net drag from impurities and irregularities in the material, causing hysteresis in the dependence of B on H . There will in general be a remanence B_r , the flux density remaining after saturation when the field is reduced to zero, and a coercive force H_c , the reverse field required to reduce the flux density to zero. A loss associated with the irreversibility of magnetization changes also occurs in rotating fields. This loss becomes zero in very large fields, except in a few special cases.

Even in very slowly changing fields, a wall characteristically moves in jumps, each giving a sudden change in B . This irregularity has been known for a long time as the Barkhausen effect, and its physical origin is the irregularity of wall motion through various inhomogeneities in the material. Usually a very large number of these small jumps takes place. In special circumstances, however, the material may remain at B_r until, in a sufficiently large field, a single wall will be nucleated and sweep all the way across the specimen, leaving it at B_r in the other direction. Such a material has only two stable states, $+B_r$ and $-B_r$, a useful behavior in some applications.

It is also necessary to consider the behavior in rapidly varying fields, discussed below.

Antiferromagnetism. Exchange forces can operate to hold neighboring moments anti-parallel, rather than parallel. Materials whose magnetic moments are arranged in this way show no external permanent moment and are called antiferromagnetic. The sign of the exchange force may depend, among other things, on the atomic spacing. Metallic manganese, for example, is antiferromagnetic, while many alloys of manganese, in which the average Mn-Mn distance is greater, are ferromagnetic. In some antiferromagnetic compounds, the exchange interaction appears to be of a next-nearest-neighbor type, taking place through an intervening atom such as oxygen. This type of interaction is termed superexchange. Antiferromagnetic materials, having no net external moment, show small positive susceptibilities that reach a maximum at the temperature above which the exchange forces can no longer hold the moments aligned against thermal agitation. This temperature, T_N , the Néel temperature, corresponds to the Curie temperature of a ferromagnet. Magnetocrystalline anisotropy exists for antiferromagnets just as for ferromagnets (see ANTIFERROMAGNETISM).

Ferrimagnetism. With more than one type of magnetic ion present, in certain compounds,

antiferromagnetic coupling may lead to a net external moment corresponding to a Bohr magneton number equal to the difference in ionic moments. Other more complicated cases occur. Ferrites, insulating oxides with the spinel structure, are important examples of this class of material, called ferrimagnetics (see FERRIMAGNETISM).

Exchange Anisotropy. A ferromagnetic phase may be in exchange coupling with an antiferromagnetic phase, as in a cobalt particle covered with CoO. This leads to new phenomena, including non-vanishing high-field rotational hysteresis. Such a material cooled in a field through the Néel temperature, if $T_c > T_N$, may exhibit a hysteresis loop that is permanently displaced from the origin. This is equivalent to a unidirectional (not uniaxial) anisotropy and will appear in a torque curve as a $\sin \theta$ term. Ferromagnetic and antiferromagnetic regions in a single-phase alloy may also lead to these effects.

Other Configurations. Atomic moments need not necessarily be either parallel or antiparallel. In a few materials they may be arranged in a triangular or spiral configurations. In some circumstances, an antiferromagnetic material may shift to a configuration having a large ferromagnetic moment in the appropriate combination of fields and temperatures (metamagnetism).

Permanent Magnets. A useful permanent magnet material should have as large a hysteresis loop as possible. In the early magnet steels, wall motion was made difficult by a heterogeneous alloy structure. A different approach is based on the theory that sufficiently small particles should find it energetically unfavorable to contain domain boundaries. The critical size is proportional to $K^{1/2}/M_s$. Reversal must then proceed by the difficult process of rotation against shape, strain-magnetostriction, or crystal anisotropy. Fine-particle ($\sim 1000\text{\AA}$) iron and iron-cobalt materials utilizing shape anisotropy have been developed. A magnetic oxide, $\text{BaO} \cdot 6\text{Fe}_2\text{O}_3$, utilizes magnetocrystalline anisotropy in fine-particle ($\sim 1\mu$) form. The Alnico permanent magnet alloys have very fine precipitate structures and are probably also best regarded as fine-particle materials.

Thin Films. Since a surface atom's surroundings are different from those in the interior, the magnetization and Curie temperature of thin films should yield important information about the range of ferromagnetic interactions. Experimental difficulties, primarily with purity, have beclouded the subject to some extent, but it now appears that any surface layer on nickel having substantially different magnetic properties from the bulk cannot be more than a few Angstroms thick.

There have been many investigations of flux reversal in films, usually vapor-deposited on glass, which have been motivated by computer technology needs. Such films show a uniaxial anisotropy associated with fields present during deposition or sometimes with geometric effects such as the angle of incidence of the vapor beam.

Dynamic Behavior of Ferromagnetic Materials. Changes in flux in a conductor induce emf's resulting in current flows whose fields tend to oppose the change in flux. For various time rates and geometries these can be calculated, leading to expressions for phase relationships and skin depth in conductors (see ELECTROMAGNETIC THEORY). These expressions have often been applied to magnetic materials at power frequencies by simply replacing H by B . This is in general not a good approximation and leads to erroneous results. It is more nearly correct to recognize that highly localized eddy currents around moving domain boundaries are the entire source of loss under these conditions. For a given dB/dt , the loss calculated in this way is much greater, decreasing to the classical value as the density of domain boundaries increases.

In bulk metals, domain wall velocities are usually determined by the damping associated with local eddy currents. In ferrites and thin films, other types of damping may predominate. These and many other aspects of the dynamic behavior of magnetic materials of all types have been investigated through resonance methods (see MAGNETIC RESONANCE).

Superparamagnetism. For particles whose volume v is on the order of 10^{-18} cm^3 or less, the direction of the entire particle moment $M_s v$ may fluctuate thermally. An assembly of such particles will exhibit the Langevin function magnetization curve of a paramagnetic with the extremely large moment $M_s v$; thus it may be easily saturated with ordinary fields and temperatures. Such magnetization curves can be used to study particle sizes and size distributions.

JOSEPH J. BECKER

References

Books

- Bozorth, R. M., "Ferromagnetism," New York, D. Van Nostrand Co., 1951.
- Kneller, E., "Ferromagnetismus," Berlin, Springer, 1962.
- Rado, G. T., and Suhl, H., Eds., "Magnetism," Vols. I and III, New York, Academic Press, 1963. Vol. II, in preparation.

Review Articles

- Kittel, C., and Galt, J. K., "Ferromagnetic Domain Theory," in Seitz, F., and Turnbull, D., Eds., "Solid State Physics," Vol. 3, New York, Academic Press, 1956.
- Becker, J. J., "Recent Developments in Magnetic Metals and Alloys," *Metallurgical Reviews*, 7, 371 (1962).
- Wohlfarth, F. P., "Hard Magnetic Materials," *Advan. Phys.* 8, 87 (1959).

MAGNETO-FLUID-MECHANICS

Magneto-fluid-mechanics is the subject that deals with the mechanics of electrically conducting fluids (such as ionized gases and liquid metals) in the presence of electric and magnetic fields.

Magneto-hydrodynamics is another name used extensively, but it suffers from the less general meaning of the words "hydro" and "dynamics." Other names used are: magneto-hydro-mechanics, magneto-gas-dynamics, magneto-plasma-dynamics, etc.

The fundamental assumptions underlying magneto-fluid-mechanics are those of continuous media. In this respect, magneto-fluid-mechanics is related to plasma physics (see PLASMAS) in the same way that ordinary fluid mechanics is related to the kinetic theory of gases. More specifically, such phenomenological coefficients as viscosity, thermal, and electrical conductivities, mass diffusivities, dielectric constant, etc., are assumed to be known functions of the thermodynamic state, as derived from microscopic considerations or experiments.

From electromagnetic theory, we know that the "Maxwell stresses" give rise to a body force made up of the following components: electrostatic (applied on a free electric space charge); ponderomotive (the macroscopic summation of the elementary Lorentz forces applied on charged particles); electrostrictive (present when the dielectric constant is a function of mass density); a force due to an inhomogeneous electric field and its magnetic counterpart; and the magnetostrictive force. For any fluid the last two forces are negligibly small at normal temperatures, whereas the ones associated with the behavior of the dielectric constant, although normally small, are of the same order of magnitude as the buoyant forces under certain conditions. On the assumption that we deal with electrically neutral but ionized fluids, the only substantial force that remains is the ponderomotive force. Indeed, what is today called magneto-fluid-mechanics deals almost exclusively with this force.

Fundamental Equations. The equations that govern magneto-fluid-mechanics are the following:

(1) Equation of conservation of mass, which is the same as in ordinary fluid mechanics.

(2) Equation of conservation of momentum, which is altered by the forces enumerated above. In particular, the ponderomotive force per unit volume is given by $\mathbf{J} \times \mathbf{B}$ where \mathbf{J} is the vector current density and \mathbf{B} the magnetic flux vector, both measured in the laboratory.

(3) Equation of energy conservation; the same as in ordinary fluid mechanics with the addition of the Joulean dissipation $\mathbf{E}' \cdot \mathbf{J}'$. The primes indicate that the electric field and current density are measured in a frame of reference moving with the fluid. In the nonrelativistic case and for zero space charge, we have $\mathbf{E}' = \mathbf{E} + \mathbf{q} \times \mathbf{B}$ and $\mathbf{J}' = \mathbf{J}$. The barycentric stream velocity is indicated by \mathbf{q} .

(4) Equation describing the thermodynamic state.

(5) Conservation of electric charge.

(6) Ampère's law.

(7) Faraday's law.

(8) Statement that the magnetic poles exist in pairs only.

(9) Ohm's phenomenological law. Equations (1) to (3) are the conservation equations. Equations (5) to (8) are Maxwell's equations. For a large number of problems, the phenomenological coefficients of electrical conductivity and the like are assumed to be scalar quantities. This implies that the collision frequencies among the particles are much higher than the cyclotron frequency associated with the tendency a charged particle has to rotate around a magnetic line under the influence of the Lorentz force. This means that the transfer of electric charge, mass, momentum, and energy is not realized in a preferential direction.

Physically, the magneto-fluid-mechanic system of equations is coupled in the following sense: A velocity field \mathbf{q} cutting magnetic lines of flux \mathbf{B} gives rise to an induced current whose magnitude is given by $\mathbf{J} = \sigma (\mathbf{q} \times \mathbf{B})$. At the same time the fluid feels an induced body force equal to $\mathbf{J} \times \mathbf{B}$. On the other hand, it is apparent that the electric currents induced by the motion create, according to Ampère's law, a magnetic flux which distorts the original applied field. The basic mechanism of this distortion is the one created by the irreversibility introduced by the finite electrical conductivity, the same way that the distortion of the inviscid streamlines in ordinary fluid mechanics takes place by the action of viscosity.

Nondimensional Parameters and Some Important Theorems. In order to study the nature of the solutions as they emerge from different problems, we shall form a number of nondimensional parameters that can be extracted from the different equations. The order of magnitude of the inertia force per unit volume is given by $\rho V^2/L$ where ρ is the mass density, V the velocity, and L a characteristic length; the order of magnitude of the ponderomotive force $\mathbf{J} \times \mathbf{B}$, after using Ohm's law, is equal to $\sigma B^2 V$. Also, the order of magnitude of the viscous force is: $\mu V/L^2$. The ratio of the typical inertia force over the viscous force is called the Reynolds number (Re) and, from the above, is found to be: $Re = \rho V L / \mu$. The ratio of the ponderomotive force over the inertia force is given by $\zeta = \sigma B^2 L / \rho V$. The ratio of the ponderomotive force over the viscous force is equal to $(Re)\zeta$ and is defined in the literature as the square of the "Hartmann number," denoted by M . We have $M = B L \sqrt{\sigma / \nu \mu}$. The distortion of the magnetic field due to the hydrodynamic field can be studied best with the help of the following two equations:

$$\frac{d\Omega}{dt} = (\Omega \cdot \nabla) \mathbf{q} + \frac{\mu}{\rho} \nabla^2 \Omega$$

$$\frac{d\mathbf{H}}{dt} = (\mathbf{H} \cdot \nabla) \mathbf{q} + \frac{1}{\sigma \mu_c} \nabla^2 \mathbf{H}$$

In the above μ_c is the magnetic permeability and \mathbf{H} the magnetic field intensity. The first equation describes the diffusion of vorticity Ω , whereas the second can be obtained by a combination of Ampère's and Ohm's laws after elimination of the electric field by using Faraday's law. In the ordinary fluid mechanic case, the streamlines obtained

after solving the inviscid problem are distorted in regions of high vorticity through the mechanism of viscosity (last term in first equation): Similarly the magnetic field calculated in the case of ideal, nondissipative flow with $\sigma = \infty$ is distorted by the finite electrical conductivity (last term in second equation). The nondimensional number describing the influence of viscosity is the Reynolds number, and in perfect analogy as indicated by the above two equations, the magnetic field distortion is described by the number, $(Re)_m = VL/(\mu_e \sigma)^{-1}$ and is called the magnetic Reynolds number. When $(Re)_m$ is zero, the magnetic lines remain undisturbed, whereas in the limit $(Re)_m \rightarrow \infty$, the magnetic lines are frozen into the fluid in exactly the same way that vorticity is frozen according to Helmholtz's theorem. Mathematical similarities apart, the freezing of the magnetic lines with the motion, is evident in the case of $\sigma \rightarrow \infty$ from the following physical considerations: An observer moving with the barycentric velocity in a medium of infinite electrical conductivity can measure only a zero electric field and hence he could not cut magnetic lines, which means that the magnetic lines must move along with his speed. From this argument it also follows that the total change in the magnetic flux through a given surface moving with the stream must be zero for an infinitely conducting medium. Finally, the remark should be made that except for stellar and interspace applications where the velocities and (especially) characteristic lengths are high, $(Re)_m$ is a small number. The assumption $(Re)_m \rightarrow 0$ is a rather drastic one since it permits the uncoupling of Maxwell's equations from the conservation equations.

For the calculation of the ponderomotive force we can use Ampère's law ($\nabla \times \mathbf{H} = \mathbf{J}$) to find that $\mathbf{J} \times \mathbf{B} = (\nabla \times \mathbf{H}) \times \mathbf{B}$. Through regular vector operations, we can prove that $(\nabla \times \mathbf{H}) \times \mathbf{B} = \text{grad} (B^2/2\mu_e) + \text{div} (\mathbf{B}\mathbf{B}/\mu_e)$. One can identify the last term as one representing a tension equal to $B^2/2\mu_e$ acting along the lines of force, whereas the first one corresponds to an equivalent hydrostatic pressure equal to $B^2/2\mu_e$. This term is frequently called "magnetic pressure" and in different problems is found to behave precisely as the static pressure does.

Consider now the propagation of small disturbances in the form of acoustic waves for which the speed of sound for an ideal gas is proportional to $\sqrt{p/\rho}$. Now one can show through a linearization of the equations of conservation, assuming the presence of the magnetic pressure alone, that a small disturbance (for a gas of infinite electrical conductivity) will be propagated, in perfect analogy, with a speed equal to $\sqrt{B^2/\rho\mu_e}$. This is the so-called Alfvén speed, and these waves are called magneto-fluid mechanic waves. Of interest also are combinations of several mechanisms of propagation which might include sound and gravitational waves. Furthermore, one can show that if the magnetic lines are lengthened, the magnetic field intensity can be increased. Because of this property of the magnetic lines, along with the additional ones of distortion and propagation of

disturbances, the properties of the magnetic field are presented in loose terms as resembling very much those of rubber bands.

Applications. There are both astrophysical and terrestrial applications of magneto-fluid-mechanics. One of the earliest ones was perhaps suggested by Faraday, who thought to harness the river Thames with electrodes on its banks that would collect the induced electric current resulting from the flow of the river as it cuts the earth's magnetic field perpendicularly. Because of the small electrical conductivity of water, the small magnetic field of the earth and the small velocities, the interaction is too weak to be useful. However, with a mutually perpendicular magnetic field, flow, and induced current density fields, a large interaction is possible in the laboratory with hot ionized gases and strong magnetic fields. This area of research is called magneto-hydrodynamic power generation, and its popularity emerges from the fact that mechanical energy can be converted to electrical without thermally stressed rotating parts. As a consequence, higher temperatures can be imparted to the working medium with better thermal efficiencies. This scheme, under development now, seems to be limited by losses due to heat transferred from the hot gas to the outside, corrosion of the electrodes, and Hall current losses. (When the gyrofrequency of the ionized particles is high compared to their collisional frequency, the particles drift in a direction parallel to the flow, and as a result, the current in the direction perpendicular to the flow, to be collected by the electrodes, diminishes. The Hall effect can be turned to some advantage if it is designed to be substantial and if the current in the direction of flow is the one to be collected.)

One of the earliest astrophysical applications of magneto-fluid-mechanics was in the area of solar physics and in particular the sunspots. Sunspots were seen and studied with the help of a telescope by Galileo about 1610. Three hundred years later, Hale discovered, through the Zeeman effect, that the magnetic field in the sunspots is very high (of the order of several thousand gauss). It was, however, in the middle 1930's and in particular after the last world war that an explanation was sought in which the magnetic field was involved. At the writing of this article, there is no complete sunspot theory. However, the majority of workers in this area agree on the following rough picture. Because of mechanical equilibrium considerations, the pressure is the same at a given distance from the center of the sun in the sunspot proper or in the photosphere which is free of a magnetic field. This means that the magnetic pressure plus the static pressure in the sunspot region must balance the static pressure in the photosphere, a fact that implies that the static pressure in the sunspot is smaller. If we picture the sunspot magnetic lines to be radial, the pressure gradient in this direction is independent of the magnetic field and balances exactly the gravitational force per unit volume pg . Hence p is constant inside and outside the sunspot. Since the static pressure is proportional to density and

temperature, the above arguments compel us to accept a lower temperature inside the spot with a resulting darkening. The only question that rises is whether the order of magnitude of the magnetic pressure is enough for the effect to be significant. This seems to be so. If we assume a magnetic field of 1500 gauss (typical in a sunspot), the magnetic pressure is about 0.1 of an atmosphere which is the typical pressure in the photosphere.

An explanation for the bipolar nature of sunspots and the difference in the sign of their polarity has been offered. The differential rotation of the sun is invoked. The toroidal magnetic lines of the sun's field lying on its surface are twisted, since for very high electrical conductivity, they are frozen with the motion. As a result, the magnetic intensity is amplified and so is the magnetic pressure. Simple considerations based on the observed kinematics of the differential rotation establish the location in latitude with time where the intensities will be high enough to give rise to sunspot activity. The result compares favorably with observations. In fact, it can be shown that the sunspot activity migrates, time-wise, from the higher latitudes towards the equator as observations show. Because the twisted field is symmetric with respect to the equatorial plane, this model describes correctly the symmetry of the activity in the north and south hemispheres along with the fact that the polarity between two symmetric sunspot pairs is opposite in sign.

Efforts have been made to discover the mechanism for the generation and maintenance of cosmic magnetic fields, such as fields in stars, the earth, and galaxies. The most promising direction seems to lie in the so called "dynamo theories." Here, some general magnetic field is assumed (not necessarily strong), which upon interaction with the motion of a conducting medium (convective, or motion due to Coriolis forces), induces currents which reinforce the original magnetic field. As the magnetic field is reinforced, the ponderomotive force suppresses the motion until some kind of a steady state for both the motion and the magnetic field is reached.

Magneto-fluid-mechanics also studies problems related to magnetic confinement of plasmas and their stability to small disturbances. Consider for instance the so-called "pinch effect." Here, a strong current is passed through a cylindrical column made up of a plasma. The axial current filaments create an azimuthal magnetic field (the magnetic lines are then rings with the cylinder axis as the locus of their centers,) and as a result, a ponderomotive force is induced which compresses the plasma radially*. Through this confinement, it is hoped that temperatures of the order of 10^6 to 10^7 K will be created so that thermonuclear fusion can take place. Such configurations are normally subject to instabilities. Consider, for instance, the case in which a small "kink" is

formed in a cylindrical plasma column, such that the rings in the concave side are pressed together, whereas the rings in the convex side are separated. As a result, the magnetic flux (and hence the magnetic pressure) will be higher on the concave side resulting in a force tending to increase the concavity. We say that this configuration is unstable, since the force induced by the imposed disturbance acts in a destabilizing direction. Note that in this configuration, the center of curvature of the undistorted plasma boundary falls inside the plasma. One can now create another example in which the curvature of the confining undistorted boundary of the plasma is opposite (the center of curvature falls in the vacuum) and show that the configuration will be stable. We can then state that a sufficient condition for stability is met when the magnetic lines are everywhere convex towards the plasma. If the magnetic lines induced by the currents going through the plasma are in an unstable configuration, externally imposed magnetic fields can be used in order to "stiffen" the configuration.

The "aurora borealis" can be explained in terms of the interaction of the solar wind (due to the continuous expansion of the solar corona with a velocity of about 500 km/sec) with the geomagnetic field. The inertia associated with this "wind" will penetrate the magnetic lines of the earth, only up to the point where the induced magnetic pressure is smaller than these inertial forces. The earth's magnetic field falls off with the inverse third power from the center of the earth. Knowing the mass density and the velocity of the solar wind, we can locate the remotest magnetic line from the earth, that is strong enough to stop the penetration of the solar corpuscles. When this happens, these particles will glide along this magnetic line and eventually will come to the foot of this line at the surface of the earth. An elementary computation shows that the latitude of this line is the one where the "aurora borealis" is observed. (see AURORA AND AIRGLOW).

Convective motions can be effectively subdued by the presence of a magnetic field. Consider, for instance, the convection in a thin horizontal layer due to heating from below. Convective cells (Bénard Cells) will be formed when the buoyant force is enough to counterbalance the viscous force of the motion. At the same time, balance of energy dictates that the heat convected upwards be equal to the heat conducted from the hot source at the bottom. The ratio of these two energies is called the "Rayleigh number," and for a given geometry, it must be higher than a critical value for the Bénard cells to appear. However, when a magnetic field is present, the ponderomotive force inhibits the motion and at the same time changes the geometry of the cell. The extent of this inhibition is given by the Hartmann number (defined earlier) so that the critical Rayleigh number is higher for higher Hartmann numbers. Available laboratory experimental results reconfirm the findings of this theory. On a cosmic scale, it has been hypothesized that the roll-like granulation in the sunspot penumbra is

* Pinch-effect devices are also useful in metallurgy where molten metals can be confined away from solid boundaries in order to remain pure.

the result of the magneto-fluid-mechanic inhibition of the motion inside regular photospheric convective cells.

Magnetic fields are also known to inhibit the onset of turbulence. For instance, consider the flow of mercury in a channel. Experiments have shown that the flow can be laminar well above the critical Reynolds number of 2000 or so, if a coil is wrapped around the pipe thus creating an axial magnetic field. The small disturbances perpendicular to the direction of the main stream will be damped out through the action of the induced retarding ponderomotive force.

Many other cosmic scale phenomena seem to be explainable through magneto-fluid-mechanics. To list but a few, there are the solar flares and filaments, the spiral structure of some galaxies, the heating of the solar corona, explosion of magnetic stars and many others. Although order-of-magnitude analyses have been suggested to explain some of these phenomena, there are no complete self-consistent theories. Such theories seem to demand a simultaneous satisfaction of all the conservation and electromagnetic equations -- a formidable, if ever possible, task. On the terrestrial scale, many applications have been undertaken, and some of them are dependent upon technological development rather than fundamental physical understanding. To give a few more examples, in addition to those already mentioned, we list magneto-fluid-mechanic liquid metal pumps, and flow meters, propulsion devices based on the acceleration of a neutral plasma through which a current and a normal magnetic field from the outside are supplied (an area called "plasma propulsion"), or a device in which positive ions (such as the ones easily produced by alkali metals) are accelerated with an electric field (ionic propulsion). Other examples are devices to reduce the heat transfer in reentry objects by using the decelerating action of a magnetic field carried by the vehicle or to use the ponderomotive force as a control force when needed for the navigation of space crafts.

PAUL S. LYKOURDIS

References

- Cowling, T. G., "Magnetohydrodynamics," New York, Interscience Publishers, 1957.
- Ferraro, V. C. A., and Plumpton, C., "An Introduction to Magneto-Fluid-Dynamics," London, Oxford University Press, 1961.
- Alfven, H., and Fulthammar, C. G., "Cosmical Electro-Dynamics," Oxford, The Clarendon Press, 1963.
- Chandrasekhar, S., "Hydrodynamic and Hydromagnetic Stability," Oxford, The Clarendon Press, 1961.

Cross-references: ASTROPHYSICS, AURORA AND AIR-GLOW, CONSERVATION LAWS AND SYMMETRY, FLUID DYNAMICS, FLUID STATICS, IONIZATION, PLASMAS; SOLAR PHYSICS.

MAGNETOHYDRODYNAMICS. See MAGNETO-FLUID DYNAMICS.

MAGNETOMETRY

The term "magnetometry" designates the scientific approaches for measuring the magnetization of materials and static magnetic fields. Magnetometers are scientific instruments used for the following purposes: (1) to measure the magnetic moment of the specimen and to determine the magnetization of materials; (2) to calibrate electromagnets and permanent magnets used in the production of magnetic fields in the laboratory, and (3) to measure the strength of magnetic fields and their components on or near the surface of the earth, as well as in space.

Four distinct principles are basic to the design of magnetometers: magnetostatic action, electromagnetic induction, deflection of carriers in semiconductors, and precession of nuclear and electronic spins. This article selects the basic features of some representative instruments from among the many available types.

The Classical Astatic Magnetometer. This is used to determine the magnetic moment of rod-shaped samples. The specimen can be exposed to the controllable homogeneous magnetic field produced in the center region of a solenoid which is appreciably longer than the specimen. Its magnetic moment is characterized by strength and direction of the field at a known distance from the sample. Two equivalent, permanent magnet needles horizontally placed and rigidly linked by a nonmagnetic rod in a vertical position comprise the measuring system. Arranged in antiparallel alignment, the needles have a wide distance between them, as compared to length. This astatic system, unaffected by the earth field, is suspended on a calibrated torsion wire. The axis of the test specimen is placed perpendicular to the axis of the lower needle and in its plane of rotation. It is possible to cancel the field of the magnetizing solenoid at the location of the sensing needle by means of a magnetically opposing solenoid located in the same plane. The magnetic moment of the specimen can be derived from the angular deviation of the needle occurring from the action of the static field. However, in order to determine the magnetization in this manner, there must be provision for homogeneous magnetic field throughout the sample volume. Therefore, the sample should be machined into a rotation ellipsoid with its long axis pointing in the direction of the needle. Calibration is by means of the well-defined moment of a coil.

With a modified astatic magnetometer, the magnetic properties of extremely small specimens in the form of fine wires or films have been measured. The specimen is arranged parallel to the axis of the astatic system and a short distance from the closely spaced short needles, in order that each needle will sense the pole of the neighboring sample. Under this method, films as thin as 10^{-5} cm, weighing less than 1 mg, have been investigated.

Concurrently, with the evolution of microwave techniques in the last two decades new magnetic materials suitable for these high frequencies have been discovered. Single crystals of ferrite, garnet,

and related substances have become significant, both technically and scientifically. As a result, new types of magnetometers have been invented to determine the magnetization of extremely small samples. The new instruments are based on the relative periodic displacement of the dipole field of the specimen against pickup coils and subsequent amplification of the small induced ac voltages.

The Vibrating Sample Magnetometer. This makes use of a mechanical oscillator that can be either a loudspeaker or a motor. The sample, attached to the lower end of a nonmagnetic shaft, moves up and down, between and parallel to the pole faces of an electromagnet. Amplitude of the oscillation is approximately 2 mm, the frequency approximately 80 cps with loudspeaker drive. The sample stays within the homogeneous region of the magnetizing field. Two series opposing signal coils with approximately 20,000 turns each are located on each side of the sample, with axes parallel to its motion and perpendicular to the exciting field.

For small samples, only dipolar field lines are linked with the signal coils. The magnetic moment can be derived directly by comparison with the voltage excited by a standard sample of known magnetization (nickel). The vibrating sample magnetometer can be operated over a wide range of temperature with a maximum sensitivity (defined as the sample magnetic size needed for a 1 to 1 signal-to-noise ratio) of 10^{-6} to 10^{-3} emu.

The Vibrating Coil Magnetometer. In this instrument, the sample is kept in a fixed position in the homogeneous region of the magnetic field between the pole faces. The signal coil oscillates at approximately 40 cps, its axis and velocity vector being collinear with the dipole axis of the specimen. The distance between coil and sample is large enough to allow for installation of temperature- and pressure-generating apparatus to include the sample. Special precautions have to be taken to eliminate the signal produced by the curvature of the magnetizing field. The measurement, which can be recorded, is continuous. The sensitivity as reported is about 10^{-2} emu.

The Pendulum Magnetometer. This is a rather simple apparatus which utilizes the ponderomotive force which a sample experiences in an inhomogeneous magnetic field. The device is based on the perception that this force is (for small deviations from the position of maximum field strength) proportional to the displacement perpendicular to the field lines between the hemispherical pole pieces of an electromagnet. This condition causes a simple harmonic motion. The specimen is fastened to a light bar whose movement is constrained only in the direction of its length and without rotation by a quinquefilar suspension. The magnetization is determined from measurements of the periods of oscillation with and without magnetic field.

The Vibrating Reed Magnetometer. This has been developed as a result of the above-described principle for obtaining a harmonic oscillation between appropriately shaped pole faces. The

pendulum is replaced by a metallic reed of nonmagnetic material, and the sample is attached to one end. The vibration of this spring is excited by a piezoelectric transducer driven from an oscillator; the resonance frequency of the reed is observed with and without field. The magnetization can be calculated by comparison with a reference sample. Accuracies of about 1 per cent can be obtained on specimens with magnetic moments of 50 emu. The vibrating reed magnetometer can also be used for measurements at elevated temperatures.

The term "magnetometer" was first applied to instruments that measured static magnetic fields. The magnetometric principle was originally developed for the study of the earth's magnetic field and its peculiarities. There is a basic identity in present-day methods used for measuring fields produced in the laboratory and the field pattern on and above the surface of the earth. Specific design of instruments and their accessories are tailored to meet particular requirements; the magnitudes of the fields to be analyzed range from approximately 10^{-5} to 10^{-6} oersted.

The "Gaussmeter" or "Fluxmeter." This is a laboratory instrument widely used for the calibration of electromagnets. It comprises a small dc generator. A small generator coil is wound on a nonmagnetic core, placed at the end of a 3 to 4 foot long axis, and driven by an ac motor with constant speed of revolution. The induced sinusoidal voltage is rectified by a commutator and read on a voltmeter calibrated in magnetic field units. The device is inherently linear; a commercial version allows full-scale readings from 2.5 to 120,000 gauss with different coils (\pm per cent accuracy). The instrument also yields an approximate determination of the direction of the field vector.

The Earth Inductor. This geomagnetic survey instrument is used for measurement of the inclination (or magnetic dip) of the earth field. A coil, connected through commutator and brushes to a sensitive galvanometer, is rotated about its diameter by a hand-powered flexible shaft. When the rotation axis of the coil is brought in line with the earth field, the galvanometer will read zero; the inclination of the axis against the horizontal plane can be read with an accuracy of approximately 0.1°.

The Classical Magnetometer of Gauss. This instrument determines the horizontal field intensity of the earth field in stationary observatories. Two measurements are required: (1) the measurement of the period of a permanent magnet, which is vertically suspended on a torsion fiber, when it oscillates in the horizontal plane about the magnetic meridian; (2) the measurement of the deflection angle, which a magnetic needle experiences attached to the same suspension, when the permanent magnet acts at a preset distance in a preferred position together with the earth field upon the needle. This method is an absolute one.

The Sine Galvanometer. This is an absolute instrument for the determination of horizontal intensity. With it, the suspended detector magnet is acted upon by a field produced by a calibrated

coil and the horizontal field component. Accurate measurements of the coil current and deflection angle of the needle are required.

The Flux-gate Magnetometer. This magnetometer, which lends itself to aircraft application, has been used successfully for detecting magnetic anomalies and for exploring the earth's surface in search of mineral deposits. It has also been used for submarine detection. The operation of this instrument is based on the change of permeability of a highly sensitive material in weak fields. A device known as the flux gate exploits this effect. It contains two permalloy cores in parallel position. A coil is wound on each of the cores; the two windings are opposed in their polarities and connected in an impedance bridge circuit in such a manner that an ac voltage supplying the bridge will not produce a diagonal voltage. The bridge is balanced in the absence of an external field. If a component of the earth's field parallels the axis of the core, the bridge becomes unbalanced due to the opposing magnetic biases. The voltage produced then acts upon servomechanisms. Measurement of the field is as follows: Three mutually perpendicular flux gates are mounted on a platform that can be rotated by servomotors in two perpendicular planes. Activated by the diagonal voltage from one flux gate, each servomotor rotates its twin core, decreasing the unbalance and holding the cores in zero position. The combined action of two gates then brings the cores of the third gate into the direction of the field. An additional field winding encloses both cores of the third gate; a servomechanism controlled by the diagonal voltage of this gate provides a current through the field winding, in order to annul the earth's field. The magnitude of this current determines the strength of the earth's field. The flux gate magnetometer is sensitive to variations of about 1 gamma ($\sim 10^{-5}$ oersted).

Hall-effect Device. In recent years, investigations on semiconductivity led to the Hall-effect device for measurement of magnetic fields. Hall voltage occurs in a current-carrying sample of semiconducting material perpendicular to the current and perpendicular to an applied static magnetic field. The magnitude of the voltage is proportional to the field. Low noise level is inherent in this device; dimensions of the sensing element can be kept rather small. The Hall voltage amounts to approximately 100 mV at 10 000 gauss, with 100 mA current. The operating temperature range is between -40 and $+85$ C for a commercial device. The use of the Hall-effect device is mainly restricted to the laboratory.

The Nuclear Precession Magnetometer. This device uses the phenomenon of nuclear magnetic resonance. This method covers a wide field range with high accuracy in determining field strengths. Nuclear probes are extensively used for determination of fields and field gradients of laboratory magnets. The technique lends itself to telemetering the information; it is capable of operating at small fields. Therefore, the proton-precession magnetometer is employed in space probes and satellites.

The nuclear resonance method is characterized

by an outstanding feature: the sharp line width for absorption of rf power by the appropriate types of nuclei when exposed to a magnetic field. The resonance frequency for H^1 is 4257.8 cps/gauss and 1654.6 cps/gauss for Li^7 . The probes used for calibrating magnets are a few millimeters in size. The resonance frequency can be determined precisely by counting procedure.

For geomagnetic measurements, a larger probe (water) is required to obtain a sufficiently strong response at small fields. First, the probe is exposed for a few minutes to a strong field (100 gauss) perpendicular to the earth's field, in order to obtain sufficient polarization. On sudden removal of this field, the spinning protons precess about the earth's field, inducing a voltage in a coil suitably placed. The signal duration is approximately one second, at which time it relaxes toward zero. Nevertheless, accurate frequency measurements can be made within this time. This total-intensity nuclear magnetometer has been improved for the determination of vector field measurement.

ERNST R. CZERLINSKY

References

- Kohlrausch, F., *Praktische Physik* 2, 80 (1943).
- Foner, S., "Vibrating Sample Magnetometer," *Rev. Sci. Instr.* 27, 548 (1956).
- Smith, D. O., "Development of a Vibrating-Coil Magnetometer," *Rev. Sci. Instr.*, 27, 261 (1956).
- Nelson, J. H., et al., "Magnetism of the Earth," *U.S. Dept. Comm. Publs.*, 40-1 (1962).
- Jensen, Homer, "The Airborne Magnetometer," *Sci. Am.*, 204, 151 (1961).
- Ingram, D. J. E., "Spectroscopy at Radio and Microwave Frequencies," pp. 102, 289, London, Butterworth, 1955.

Cross-references: GEOPHYSICS; HALL EFFECT AND RELATED PHENOMENA; MAGNETISM; MEASUREMENTS, PRINCIPLES OF.

MAGNETOSTRICTION

When a polycrystalline nickel sample is placed in a magnetic field, it contracts along the field direction by about 30 parts per million and elongates in the transverse direction by about half that amount. There is also a small volume change. Such changes in dimension of magnetic materials with variation of magnetic field strength or direction, are termed *magnetostriction*. They are measured by strain gages, optical dilatometers, capacitance variation, and x-ray analysis.

Below the Curie temperature, magnetostriction in weak fields is determined by domain rotation, becoming appreciable at fields near the knee of the *B-H* curve.

In saturating fields there is still a small linear dependence of magnetostriction on magnetic field strength, and above the magnetic ordering temperature magnetostriction is, except in rare instances, quadratic in magnetic field strength.

Field strength dependent distortions in the saturated and paramagnetic regions, designated *forced magnetostriction*, are due to the paraprocess, the induction of a moment by the field.

The saturation magnetostriction of single crystals depends upon the direction of the (sublattice) magnetization, α , and the direction of measurement, β , with respect to the crystal axes. In a cubic crystal (with collinear sublattices), to lowest order,

$$\delta l = \lambda_0 + \frac{1}{2}\lambda_{100}[\alpha_1^2\beta_1^2 + \alpha_2^2\beta_2^2 + \alpha_3^2\beta_3^2] + 3\lambda_{111}[\alpha_1\alpha_2\beta_1\beta_2 + \alpha_2\alpha_3\beta_2\beta_3 + \alpha_3\alpha_1\beta_3\beta_1] \quad (1)$$

Birss¹ gives higher order expressions for cubic and hexagonal symmetry and references for other symmetries. The distortion of an unmagnetized polycrystalline material placed in a saturating magnetic field can be calculated by averaging over directions in Eq. (1). For cubic polycrystals the change in length parallel to the field accompanying magnetization is given by

$$\bar{\lambda}_H = \frac{2\lambda_{100} + 3\lambda_{111}}{5} \quad (2)$$

Birss¹ emphasizes the unreliability of the assumption of initially random domain distribution implicit in Eq. (2).

Magnetostriction coefficients vary greatly, depending upon the material, temperature, and magnetization state. For pure iron at room temperature, the saturation magnetostriction constants are $\lambda_{100} \sim 20 \cdot 10^{-6}$; $\lambda_{111} \sim 20 \cdot 10^{-6}$, while for alloys near 80Ni-20Fe (weight per cent) these constants are almost zero. The cobalt ion causes a large magnetostriction; for cobalt ferrite $\lambda_{100} \sim -500 \cdot 10^{-6}$ while for nickel ferrite $\lambda_{100} \sim -30 \cdot 10^{-6}$. The largest known magnetostriction is that of dysprosium metal.² As a magnetic field is rotated in the basal plane of this hexagonal crystal, there is a basal plane distortion of almost one per cent, at liquid nitrogen temperatures and below.

The source of magnetostriction is the dependence of magnetic energy on strain. Because the elastic energy is quadratic in strain while the magnetoelastic energy is linear in strain, the minimum free energy occurs at nonzero strain. For example, in a cubic crystal the equilibrium shear strain ϵ_{xy} is given by

$$\epsilon_{xy} = \frac{B_2(T, H)}{c_{44}} \alpha_x \alpha_y \quad (3)$$

Here c_{44} is the elastic constant, the α 's are magnetization direction cosines, and $B_2(T, H)$ is a magnetoelastic coefficient representing the variation of magnetic energy (magnetic anisotropy, dipolar, anisotropic exchange) with strain.

Quantum mechanical calculations of the magnetoelastic coefficients are in a somewhat more satisfactory state in the case of nonconductors than for metals. Extensive calculations by

Tsuya of the B coefficients of the spinels are reviewed by Kanamori.³

The temperature dependence (and "forced" field dependence) of the magnetostriction coefficients is due to statistical averaging as the individual spins fluctuate around the average magnetization direction α . For some materials this temperature dependence can be expressed entirely in terms of a known function of the (sublattice) magnetization. For ferrimagnets⁴

$$\lambda_i(T, H) = \sum_n \lambda_i^n(0) f_i(m_n(T, H)) \quad (4)$$

That is, the magnetostriction coefficient $\lambda_i(T, H)$ is the sum over sublattices of temperature independent sublattice magnetostriction coefficients ($\lambda_i^n(0) - B_2^n(0)/c_{11}$) times a function f_i of the sublattice magnetization $M_n(T, H)/M_n(0)$. At sufficiently low temperatures this function reduces to

$$f(m_n(T, H)) = \left[\frac{M_n(T, H)}{M_n(0)} \right]^3; T \ll T_c \quad (5)$$

for both λ_{100} and λ_{111} .

Bozorth,⁵ Carr,⁶ and Kneller⁷ give references on magnetostriction and discuss measurement techniques.

EARL CALLEN

References

1. Birss, R. R., *Advan. Phys.*, **8**, 252 (1959)
2. Legvold, S., Alstad, J., Rhyne, J., *Phys. Rev. Letters* **10**, 509 (1963); Clark, A. E., Bozorth, R. M., and DeSavage, B., *Physics Letters*, **5**, 100 (1963).
3. Kanamori, J., "Magnetism," (Rado, G. T., and Suhl, H., Eds.), Vol. I, p. 127, New York, Academic Press, 1963.
4. Callen, E., Clark, A. E., DeSavage, B., Coleman, W., and Callen, H. B., *Phys. Rev.*, **130**, 1735 (1963).
5. Bozorth, R. M., *Ferrromagnetism*, New York, D. Van Nostrand, 1951.
6. Carr, W. J., Jr., "Magnetic Properties of Metals and Alloys," Cleveland, Ohio, American Society for Metals, 1959.
7. Kneller, E., "Ferromagnetismus," Berlin, Springer, 1962.

MANY-BODY PROBLEM

Scope and Definition. A large part of the experimental data of physics is concerned with natural objects which may be looked upon as being made up from smaller bodies. For example, we may think of the solar system as an object composed of the planets and the sun; ordinary matter, in solid, liquid or gaseous form, as composed of molecules and atoms; atoms and molecules themselves as made up from nuclei and electrons; and finally, the nuclei as composed from neutrons and protons. We shall call the composite object the system, and its constituents, the particles; and note that it seems most reasonable to suppose that the properties of the system can be explained on the basis of the law of interaction between the particles and the laws of dynamics.

The latter may be classical or quantum mechanical according to the demands of the situation. At each level of refinement we refrain from asking about the internal structure of the particles. This is to achieve a natural simplicity of description; but still, at each such level, we have a rich variety of natural phenomena to explain.

The many-body theory is not concerned with any fundamental or complete explanation of nature. Its chief aim is to formulate schemes according to which calculations of certain physical quantities can be performed theoretically and the results can be compared with experimental measurements. It is inherent in its methods that the number of particles is considered as being large, and no attempt is made to find all the details of motions of the particles—a characteristic which distinguishes it from the so-called one-, two- or three-body problems.

Since the early 1950's there has been a great deal of activity in the theory of quantum mechanical many-body systems such as nuclei, solids and fluids. As a result, the term many-body problem has come to mean, almost exclusively, the theory of such systems at or near the absolute zero of temperature. The latter qualification serves to distinguish the many-body problem as such from the closely related field of STATISTICAL MECHANICS. These modern developments are based on the observation that when the number of particles is so large that it may be considered effectively infinite, then the system becomes very similar to that of interacting fields—except for the nature of interactions considered and the general formal methods of quantum field theory and quantum electrodynamics may be used with advantage.

There is only one general theorem in many-body theory; it is known as Poincaré's theorem. Roughly speaking, it states that any given initial state of a finite many-body system will be repeated provided one waits long enough. The quantum mechanical form of this theorem states that all observables in a finite system are almost periodic functions of time. This theorem has not had much practical use but has played an important role in discussions concerning the foundations of statistical mechanics.

The so-called many-body theory is mainly a collection of special approximate methods developed for particular problems. The chief common features of some of the methods, especially the ones connected with modern developments, will now be described.

Reduction to an Equivalent System of Non-interacting Particles. The very fact that we can recognize the constituent particles leads us to believe that in the lowest approximation we may neglect their interactions. This approximation is already quite successful in derivation of perfect gas laws and electron theory of metals. A slightly different form of this assumption occurs in the case of atoms, which are treated as systems of non-interacting electrons moving in the field of force of the nucleus. For planetary systems a similar approximation is used.

The normal mode analysis of a lattice provides an example where a transformation of coordinates is used to achieve such a reduction. Instead of considering the coordinates of individual particles which interact with each other through harmonic forces, one considers certain linear combinations of displacements, the modes. In terms of the new variables there are no interactions and the solution is immediately obtained. This is an example of a *transformation* which introduces a *collective* description of the system.

Another type of situation occurs in nuclear theory, where it is found that a shell-model of the nucleus, built in analogy with the atomic shell-model, is very successful. The non-interacting particles of this model are called neutrons and protons, but interaction between them which must be used in this model is vastly different from that observed in two-body scattering experiments. As a first approximation one can completely ignore the mutual interaction and assume that the particles move in a common one-body potential. This circumstance suggests that what are called neutrons and protons in this model are not the same as the free ones but are only some *quasi-particles* which are appropriate to the model and happen to have many properties in common with actual particles. An analogous situation occurs in some solids where the electrons are observed, by means of cyclotron resonance experiments, to possess an effective mass different from the mass of free electrons.

Effective Field Method. This is one of the methods of taking into account the mutual interactions of the particles. One starts with a given motion of particles, e.g., from an approximation of the type described in the last paragraph, and calculates the field of force experienced by one of the particles under the influence of all the others. As a further refinement the field may be made *self-consistent*, that being the situation when the motion produced under the influence of the field is the same as that which generated it. But for the approximations made in the course of calculation, such as omission of the effects of correlations among the particles, a fully self-consistent theory would be a complete theory.

Examples are: Hartree-Fock theory, Fermi-Thomas approximations, Brueckner theory of nuclei, Wigner-Seitz cell model in solid state theory, and several others.

Collective Motion Theory. In some phenomena, such as propagation of sound and plasma oscillations, it is clear that many particles are performing coordinated movements. To study such cases, one introduces some collective variables in addition to the usual ones, and the Hamiltonian is re-expressed in terms of these mixed variables. *Subsidiary conditions* have to be imposed upon this extended system of variables to preserve the original number of degrees of freedom. The collective variables should be such that there is no appreciable interaction between these and other degrees of freedom. When quantum mechanics is applicable, the collective motions are also excited in quanta which for all practical purposes may be

treated as new (quasi) particles. The stability of collective motions is then expressed in terms of the lifetimes of quasiparticles. Solid-state physics is particularly rich in exhibiting collective motions. Quasiparticles associated with some of them are: the phonons (sound, lattice vibration); the polarons (electron and its polarization field in dielectric); and the excitons (electron-hole excitations in insulators). Collective motions in nuclei can also be interpreted in a similar manner. Superconductivity and superfluidity are also examples of collective motion. The quasiparticles responsible for superconductivity are electron pairs with equal and opposite momenta.

Use of Techniques of Field Theory. With these techniques it is possible to obtain formal expressions which represent the effect of interparticle interactions to any order in perturbation theory. By carrying out rearrangements and partial summations of terms in perturbation series it is possible to see that, as far as the motion inside the system is concerned, the relationship between the coordinates and momenta and the potential and kinetic energies is changed in such a way that it has to be described in terms of an effective mass and an effective interaction, which differ from the original quantities in a known way. In certain cases these effects can be calculated and are finite.

Brueckner's theory of nuclear matter is an example of this type. The effective mass is found to depend on the momentum of the particle inside the system, and the effective interaction, the so-called t - or K -matrix, is given by an integral equation involving the original interaction. A self-consistent calculation of the properties of the system (nuclei or atoms) can be based on this understanding.

Similar techniques can be used for studying collective motions. An example is the treatment of electron gas by Gell-Mann and Brueckner.

Perhaps the greatest advance has been made in the theory of superconductivity, where variational and canonical transformation methods have been used.

A combination of all these methods is needed to study the difficult problem of relationship between various excitations, i.e., the interaction between various quasiparticles of a many-body system. One of the most useful tools in these calculations is the representation of matrix elements by means of diagrams, first introduced by Feynman. Many of these methods were first developed in connection with the theory of interacting fields and they are usually employed in many-body problems for the limiting case of an infinite number of particles, but these restrictions are not essential; in fact, they are quite general methods of treating arbitrary quantum mechanical systems.

KAILASH KUMAR

References

- ter Haar, D., "Introduction to the Physics of Many-Body Systems," New York, Interscience Publishers, 1958.
- De Witt, B., "The Many-Body Problem," London, Methuen, 1959.
- Thouless, D. J., "The Quantum Mechanics of Many-Body Systems," New York, Academic Press, 1961.
- Kumar, K., "Perturbation Theory and the Nuclear Many-Body Problem," Amsterdam, North-Holland Publishing Co., 1962.
- Khilm, G. F., "Qualitative Methods in Many-Body Problem," New York, Gordon and Breach, 1961, (for classical mechanics).
- Collections of reprints of original articles and lecture notes on many-body problem have been published by various firms, e.g., Pines, D., "The Many-Body Problem," and Van Hove, L., Hugenholtz, N. M., and Howland, L. P., "Quantum Theory of Many-Particle Systems," New York, W. A. Benjamin, 1961.
- Cross-references:** EXCITON, FIELD THEORY, KINETIC THEORY, MATRIX MECHANICS, NUCLEAR STRUCTURE, PHONON, PLASMAS, QUANTUM THEORY, QUANTUM ELECTRODYNAMICS, SOLID STATE THEORY, STATISTICAL MECHANICS, SUPERCONDUCTIVITY, SUPERFLUIDITY.

MASER

The term "maser," coined by Townes and co-workers who pioneered this field, stands for *m*(icrowave) *a*(mplification by) *s*(timulated) *e*(mission of) *r*(adiation). "Microwave" has proved restrictive; stimulated emission amplifiers have operated in the UHF (~ 300 mc/sec), and at infrared, visible, and ultraviolet frequencies (see LASER). The principal advantage of the maser amplifier is its small intrinsic internal noise: the equivalent *noise input temperature* is but a few degrees Kelvin. The theoretical minimum noise input temperature is hf/k , where h is Planck's constant, k is Boltzmann's constant, and f is the signal frequency. This is 0.48 K at $f = 10$ Gc/sec (gigacycles per second) or 10×10^9 cps. Maser oscillators can generate exceedingly monochromatic radiation, (e.g., the ammonia maser has a short-term frequency stability of ~ 5 parts in 10^{12}), and may prove to be the only practicable sources of coherent sub-millimeter wavelength radiation.

Stimulated Emission of Radiation. Because its energy is quantized, a molecule (here a generic term) can exchange energy with the electromagnetic radiation field only in discrete amounts (quanta). The emission or absorption of a quantum (photon) is associated with a transition between molecular energy states. For two states, $|m\rangle$, $|n\rangle$ of energies W_m , W_n , ($W_m > W_n$), the frequency f_{mn} of the radiation accompanying the (permitted) transition between them satisfies the Bohr condition

$$hf_{mn} = W_m - W_n \quad (1)$$

A molecule in state $|n\rangle$, exposed to radiation of frequency f_{mn} and energy density u , has a probability per unit time $u \times B_{nm}$ (B_{nm} is a constant) of absorbing a photon hf_{mn} and reaching state $|m\rangle$. There is also a probability $u \times B_{mn}$ that a molecule in the upper state $|m\rangle$ will *emit*

a photon hf_{mn} and return to the lower state $|n\rangle$. The upper state molecule is *stimulated* to emit radiation of frequency f_{mn} by the radiation field at this frequency. Stimulated emission, like absorption, is a process which is *phase coherent* with the incident radiation. Thermodynamical arguments by Einstein (1917) showed that

$$B_{nm} = B_{mn} \quad (2)$$

A molecule in the upper energy state $|m\rangle$ may also revert to the lower state $|n\rangle$ by *spontaneously* emitting radiation of frequency f_{mn} . This spontaneous emission is a random process, which is phase incoherent with any incident radiation and is therefore a source of noise in a maser.

The spontaneous emission probability A_{mn} is given by

$$A_{mn} = B_{mn} \cdot hf_{mn} \cdot \rho_f \quad (3)$$

where ρ_f is the number of wave modes per unit volume per unit frequency range open to radiation of frequency f_{mn} . Table I shows values of ρ_f under various conditions; c is the velocity of light, v_g is the group velocity of radiation.

TABLE I

Environment	ρ_f
Enclosure large compared with the wavelength c/f_{mn}	$8\pi f_{mn}^2/c^3$
Single mode resonant cavity, volume V , width of half-power response Δf	$2/(\Delta f \cdot \pi V)$
Waveguide, cross section A	$1/Av_g$

In the microwave region (say, 1 to 100 Gc/sec), $A_{mn} \ll B_{mn}$; spontaneous emission is therefore negligible except as a source of noise. However, maser spontaneous emission noise is usually exceeded by noise arising from losses in ancillary microwave circuit elements.

Molecular transitions are excited by either the electric or magnetic component of the radiation field, depending upon whether the change in molecular energy is primarily electric or magnetic in character. Each radiative transition has associated with it an effective oscillating electric or magnetic moment, usually dipolar. The probability B_{mn} given above depends directly on this dipole moment and inversely on the frequency spread (line width) δ of the transition.

Conditions for Amplification. Suppose radiation of frequency f_{mn} is incident on an assembly of molecules with an allowed transition at this frequency [Eq. (1)]. Let the number of molecules in the upper state $|m\rangle$ be N_m , and in the lower state $|n\rangle$ be N_n . If the incident radiation energy density is u , the power absorbed by the molecules will be

$$P_A = N_n u B_{mn} h f_{mn} \quad (4)$$

and the power emitted will be (see equation 2)

$$P_E = N_m u B_{mn} h f_{mn} \quad (5)$$

Since at microwave frequencies spontaneous emission is negligible, the condition for amplification is

$$P_E > P_A; \text{ i.e., } N_m > N_n \quad (6)$$

There must be an excess of molecules in the *upper* energy state of the transition associated with the signal frequency.

For thermal equilibrium at temperature T , Boltzmann statistics give

$$(N_m/N_n) = \exp(- (W_m - W_n)/kT) \\ \exp(-hf_{mn}/kT) \approx 1 - (hf_{mn}/kT) \quad (7)$$

at microwave frequencies, where $hf \ll kT$. Clearly a molecular system in thermal equilibrium is thus always absorptive. Equation (7) allows the definition of an "effective temperature" T_m for an emissive system; Eq. (6) and (7) show that T_m will be a "negative" temperature, and that $|T_m| \rightarrow 0$ for $(N_m/N_n) \rightarrow \infty$. Obtaining an emissive condition, obtaining a "negative temperature," and obtaining "population inversion" are thus synonymous. The excitation of a molecular assembly to an emissive condition is perhaps the crux of the maser problem. The schemes used depend on the conditions and on the molecular system. Discontinuous methods (pulse inversion, adiabatic fast passage) can be used, but the account here is confined to the principles of continuous methods. In a gas, actual separation of the upper-state molecules may be possible. For example, the upper-state molecules for the 23.87-Gc/sec ammonia maser transition tend to increase their energy in a static electric field, while the lower-state molecules tend to decrease their energy (Quadratic Stark effect). In an inhomogeneous electric field, the wanted upper-state molecules will therefore drift to the low-field regions. An electrode system (with geometrical axial symmetry) which gives a low-field region along the symmetry axis will therefore confine the upper-state molecules in a beam along this axis while rejecting the lower-state ones.

Most masers operate on the multilevel excitation scheme, requiring an input of energy ("pumping") at some frequency other than the transition frequency; forms of energy other than electromagnetic may also be used. The principles of the scheme will be illustrated by reference to a molecule having 3 levels with energies $W_1 < W_2 < W_3$, such that all transitions between levels are allowed. (The transitions other than the signal transition need not radiate electromagnetically). In thermal equilibrium the number densities $(n_i)_e$ of the particles in the different states (*i*) will satisfy

$$(n_1)_e > (n_2)_e > (n_3)_e$$

The frequencies f_{32}, f_{21}, f_{31} are defined from

$$f_{mn} = (W_m - W_n)/h$$

Suppose now by some means, that the transition $1 \rightarrow 3$ is *saturated*, i.e., $n_1 \approx n_3$. (This might be achieved by a sufficiently strong electromagnetic field at frequency f_{31} —known as the "pump")

frequency). Under these conditions, it may happen either that $n_2 > n_1$, or that $n_3 > n_2$. In the first case, amplification will be possible at f_{21} ; in the second case, at f_{32} , provided that the appropriate transition is electromagnetically radiative.

There are many variants on the simple scheme just described. The frequency f_{31} may lie in the optical region (OPTICAL PUMPING); the excitation may be by collision processes in a gas discharge; or more than three levels may be involved, and pump frequencies lower than the signal frequency can sometimes be used.

Maser Materials. Maser action has been achieved in gases (e.g., ammonia, formaldehyde, hydrogen, rubidium vapor) and liquids (e.g., protons in water) but the most important maser materials are the solid-state ones, since these have a high concentration of active centers in a small space. Present emphasis is on the use of certain paramagnetic ions diluted in a host crystal lattice. Three-level excitation, or some variant, is usually employed.

PARAMAGNETISM is associated with ELECTRON SPIN. The directional quantization of angular momentum leads to the quantization of the energy of the ionic magnetic moments in a steady magnetic field. In general, the ground-state multiplet of these ions is split by the crystal field of the host lattice (Stark effect), and the levels are completely separated by steady magnetic field (Zeeman effect). When the steady magnetic field is applied at an angle to the major symmetry axis of the crystal field, and the resultant Zeeman splitting is comparable with the initial Stark splitting of the levels, the usually forbidden "leap-frog" transitions necessary for 3- or multiple-level excitation become allowed. In crystal fields of low symmetry, "leap-frog" transitions may be allowed at very low or even zero magnetic fields. Clearly, ions having three or more energy levels are wanted, and any processes competing with radiative processes—e.g., the interaction of the "spins" with the lattice—are usually required to be small. Spin-lattice interaction can usually be reduced by cooling the lattice to a low absolute temperature; and indeed most solid-state paramagnetic masers operate at liquid nitrogen (77°K) or liquid helium (4.2°K) temperatures. Some ions and host lattices with which maser action has been achieved are listed in Table 2.

TABLE 2

Ion	Effective Spin	Host Lattice
Cr^{3+}	3/2	Al_2O_3 , alumina (ruby)
Cr^{3+}	3/2	TiO_2 , rutile
Fe^{3+}	5/2	Al_2O_3 , alumina
Fe^{3+}	5/2	TiO_2 , rutile
Gd^{3+}	7/2	$\text{La}(\text{C}_2\text{H}_3\text{SO}_4)_3 \cdot 9\text{H}_2\text{O}$ (lanthanum ethylsulfate)

A "spin-spin" interaction process, known as *cross-relaxation* must also be taken into account,

as it may either aid or inhibit maser action. Cross relaxation is dependent on spin concentration, but not on temperature. Consequently, maser action may be achieved at comparatively high temperature (77°K) but not at low temperature (4.2°K) where the considerably longer spin-lattice relaxation time might be expected to give better maser action. Rearrangement of the level populations occurs because of single or multiple quantum transitions between the levels, in which energy is "almost" conserved on the microscopic scale, any differential being exchanged with the energy of the macroscopic spin system (total magnetic moment).

Amplifier Systems. Maser amplifiers may be of either traveling-wave or resonant circuit (cavity) form. Their performances are expressed in terms of a molecular Q -factor, Q_m , defined over unit length for the traveling-wave maser and over the resonator volume for a cavity maser. At the signal frequency f ,

$$Q_m = 2\pi f \cdot \frac{\text{Energy stored in the structure}}{\text{Power emitted by the molecules}} \quad (8)$$

since the Q 's similarly defined for losses are positive.

For a magnetic dipole transition,

$$|Q_m| \propto \delta(N^*p_m^2\eta)^{-1} \quad (9)$$

where δ is the frequency width of the transition at half-intensity, N^* is the *excess* upper level population, p_m is the effective dipole moment for the transition, and η is the ratio of the magnetic energy coupled to the molecules to that stored in the microwave circuit.

Traveling-wave Maser. The active maser material is placed in a waveguide carrying a pure traveling wave. The gain coefficient α_m is defined such that the power gain G for a length l of amplifier is given by

$$G = \exp(2\alpha_m l) \quad (10)$$

It can be shown that

$$\alpha_m = (2\pi f_s) / (|Q_m| v_g)^{-1} \quad (11)$$

where v_g is the group velocity of radiation in the guide. Because p_m is typically of the order of a Bohr magneton, and the active centers are diluted, it is necessary to use *slow-wave structures* ($v_g \approx c/100$) in order to keep l to a reasonable value (a few centimeters). Suitable values of v_g are readily achieved by the resonant slowing obtained in periodic structures. Systems such as the Karp structure, comb structure, and meander line are favored, since these support waves with the magnetic field circularly polarized in a plane containing the direction of propagation and perpendicular to the plane of the periodic elements. The sense of circular polarization is reversed on crossing this plane and is opposite in any reflected wave to that in the forward wave. The nonreciprocal gyromagnetic properties of para- and

ferrimagnetic materials may then be employed to obtain forward gain and reverse attenuation with these slow waveguides.

The noise input temperature T_{in} of a traveling wave maser is given approximately by

$$T_{in} \simeq |T_m| + T_1(|Q_m|/Q_e) \quad (12)$$

where T_m is the effective negative temperature of the maser material, Q_m is the molecular Q (negative), Q_e is the similarly defined ohmic loss factor, and T_1 is the actual temperature of the waveguide (and contents). In this approximation, $|Q_m| \ll Q_e$. The bandwidth b_m of the amplifier is approximately equal to, but less than, δ .

The *Resonant circuit Maser* may be of either transmission (two-port) or reflection (one-port) type; only the reflection type is considered here, since it is superior in performance to the transmission type. Assuming that the unloaded resonant circuit (cavity) losses are negligible, the coupling to the external circuits will give rise to a Q -factor Q_e , say. The power gain G of the reflection cavity maser is then given by

$$G = (Q_e + |Q_m|)^2 / (Q_e - |Q_m|)^2 \quad (13)$$

The bandwidth b_e depends on the gain in such a way that

$$G^{1/2} b_e \simeq 2|Q_m| f_0 \text{ (for } G \gg 10, \text{ say)}$$

The noise input temperature is given by Eq. (12) above, where now

$$Q_e^{-1} = Q_e^{-1} + |Q_m|^{-1}$$

It is necessary to have some nonreciprocal device to separate the reflected amplified output from the input signal; the ferrite circulator is most commonly used. The bandwidth and gain stability of the cavity maser are inferior to that of the traveling-wave maser, but the cavity maser is more easily constructed. If three-level excitation is used, it is clear that any maser system must support both "pump" and signal frequencies.

Maser Oscillators. Equation (13) indicates that if $|Q_m|$ is small enough, G becomes infinite; i.e., oscillation occurs when the stimulated emission is small enough to overcome all losses. The width of the signal emitted by a maser oscillator is very much less than δ , so that for narrow δ an extremely pure oscillation signal results, and a molecular transition which is relatively insensitive to external influences will thus give oscillations of high stability in frequency. The ammonia maser and the atomic hydrogen maser are two examples.

Applications. Maser amplifiers are now in use wherever the requirement for a very low noise amplifier outweighs the technological problems of cooling to low temperatures. They have been used in passive and active radioastronomical work, in satellite communications ("Project Echo") and as preamplifiers for microwave spectrometry. The ammonia and the atomic hydrogen masers are being studied as frequency standards and have been used in a new accurate test of special relativity. Sources and

amplifiers in the submillimeter, micron, and optical wavelength regions are being studied and developed (see LASER).

G. J. TROUP

References

Review Articles

- Weber, J., *Rev. Mod. Phys.*, **31**, 681 (1959).
Wittke, J. P., *Proc. Inst. Radio Engrs. (N. Y.)*, **45**, 291 (1957).

Books

- Brotherton, M., "Masers and Lasers," New York, McGraw-Hill Book Co., 1964.
Siegman, A. E., "Microwave Solid State Masers," New York, McGraw-Hill Book Co., 1964.
Singer, J., "Masers," New York, John Wiley & Sons, 1959.
Troup, G., "Masers and Lasers," Second edition, London, Methuen and Co., 1963.
Vuylsteke, A. A., "Elementary Maser Theory," Van Nostrand, 1960.

Cross-references: ELECTRON SPIN, FERRIMAGNETISM, LASER, LIGHT, OPTICAL PUMPING, PARAMAGNETISM, ZEEMAN AND STARK EFFECTS.

MASS. See STATICS.

MASS AND INERTIA

One of the most fundamental and earliest known of physical phenomena is the simple fact that it takes some effort to push any object. To set some objects into motion by pushing is easier than to do the same for others. This property by virtue of which every body, however small, requires some force to push it is called "inertia."

As can easily be seen, the property of inertia is more general. Not only does it take some force to set a body in motion or to speed it up, but also to slow it down or stop it. The fact that a ball rolled on the ground soon stops is only because external forces, such as the friction on the ground, work on it to stop it. In these space days, it is not hard to believe that if a body were moving in empty space with no friction or air resistance, it would continue to move at the same speed without stopping. We also know that it takes force even to change the direction of motion of an object without any change of speed. One can feel this force while taking a car around a curve even when there is no accompanying change of speed.

All these are different manifestations of the property of inertia. This was neatly summed up by Sir Isaac Newton in the first of his laws of motion, which essentially says that a body will continue in its state of rest or of uniform motion along a straight line unless acted upon by an external force.

Given that it takes some force to change the state of motion of a body, the next question is—how much? Or, in other words, how do we

quantitatively measure the inertia of a body? Experience tells us that "heavier" or "more massive" bodies have more inertia. The exact technical measure of inertia is "mass," which is closely related to "heaviness" or "weight." The larger the mass of an object, the more the force it takes to change its motion by a given amount, i.e., to give it a certain acceleration. Newton's second law of motion tells us that the force required is just the product of the mass of the body times the acceleration given to it.

Force = mass \times acceleration

During the early stages of the development of the concept of mass, attempts were made to define it, not primarily as an index of inertia but as the "quantity of matter" in the body. Inertia was then considered to be a consequence of, and proportional to, this mass. However, it is not so easy to define precisely what one means by the "amount of matter" in an object. Clearly, one cannot use the size or volume of an object as an index of the quantity of matter or mass since a ball of wool and a ball of steel of the same size do not have the same mass. The "weight" of a body is quite often used to measure the mass, but this again is unsatisfactory since the weight of the same object can vary from place to place, even on different parts of our earth. Thus, it is best to understand mass as primarily an index of inertia, to which such properties as weight, size, etc., are closely related.

The fact that weight is such a good index of mass and has so successfully been used as a measure of mass is the key to some very important advances and speculations in physics. No discussion of inertia is complete without referring to this aspect of the story.

As we mentioned before (Newton's second law), under the action of a given force, e.g., a certain push of the hand, a body of larger mass M accelerates less than one of smaller mass m . This is evident repeatedly in daily life. However, there is one particular type of force, i.e. the gravitational force, under which all bodies react the same way. A "heavy" body and a "light" body, when dropped from a height reach the ground at the same time (except for small air viscosity corrections). The reason for this, well known to Newton himself, is that the gravitational force on a body, unlike any other type of force, is proportional to the mass or "inertia" of the body. Thus, a steel ball has large inertia and requires a large force to accelerate it by a certain amount. But the gravitational force on the steel ball is also correspondingly larger, so that under the influence of gravity alone, it would move exactly the same way a ball of cotton would. The fact that weight, which is essentially a name for the force of gravity, has been used as an index of mass is also a result of this proportionality between the two quantities.

All this, as we said, has been well known since the seventeenth century, but the deep significance of this apparent coincidence lay hidden until the time of the Viennese philosopher Ernst Mach

(1838–1916). Mach's views, developed, modified and put on a firm mathematical footing by Einstein, form one of the cornerstones of the latter's brilliant general theory of relativity. Giving a wide berth to the complexities of the general theory, we will only mention here that according to general relativists, the inertia possessed by a body is a consequence of the gravitational force acting upon it from all the stars and galaxies in the universe. When we attempt to push an object, we are accelerating it relative to all the distant massive fixed stars. This produces a resultant gravitational force, resisting the acceleration we are trying to give. This is why we have to exert a force to push a body.

Much of this is speculative and is truly meaningful only in the mathematical framework of the theory. However, interested readers may find a very lucid and simple discussion in the book "The Unity of the Universe" by D. W. Sciama.*

R. RAJARAMAN

Cross-references: DYNAMICS, FRICTION, MECHANICS, RELATIVITY, STATICS.

MASS SPECTROMETRY

Mass spectrometry is based on observations of the behavior of positive rays by Thompson and Wien. In 1919, Aston demonstrated the existence of isotopes by introducing neon gas into a mass spectrograph. Prior to 1940, mass spectrographs and spectrometers were used primarily for isotopic studies in university laboratories. Analytical spectrometers became commercially available during the early years of World War II when their use for the rapid analysis of hydrocarbon mixtures was recognized.

Mass spectrometry provides information concerning the mass-to-charge ratio and the abundance of positive ions produced from gaseous species. There are several techniques for the production and measurement of the ions, and the design of an instrument is determined by its proposed application. The mass spectrograph, using a photographic plate for ion detection was used primarily for isotopic studies but is now finding wide application for the analysis of trace constituents in solids. The mass spectrometer uses an electrical detection and recording system giving a metered output which provides a more accurate measure of the abundance of the ions than the photographic plate. The mass spectrometer is used primarily for the quantitative analysis of gases, liquids, and a limited number of solids.

The five basic components of the instrument are the sample introduction system, the ion source, the mass analyzer, the ion detector, and the recorder. A sample pressure of approximately 5×10^{-5} torr is generally required for a satisfactory analysis. An elevated temperature inlet system or other means of converting the sample

* Doubleday and Co., Garden City, N.Y.

into a gaseous state is required for less volatile species.

The most common methods of producing positive ions are electron impact, thermal ionization, spark and arc sources, and field emission. The electron impact source is the most widely used. Positive ions are produced by removing one or more electrons from the molecules. Thermal ionization produces positive ions by vaporizing material directly into the ion source from a filament coated with the sample. With a spark source, the material under investigation must be a conductor or else suitable means must be provided for initiating and maintaining a spark. Ions produced in the spark are taken directly into the mass analyzer. In the field emission source, a high potential is applied between the sample—generally deposited on the tip of a tungsten wire—and another electrode commonly in the form of a ring. Ionic species representative of the sample are removed by a high-intensity electric field.

The three most widely used types of mass spectrometers are (1) the single-focusing, magnetic deflection, (2) the double-focusing and (3) the time-of-flight (T.O.F.). These three types of instruments differ primarily in the method used for mass separation. The single-focusing analyzer with magnetic deflection is the most common design. The single-focusing analyzer achieves direction but not velocity focusing of the ions. Ions of the same mass-to-charge ratio, having slightly different velocities resulting from different kinetic energies imparted in the ionization process, will not be focused simultaneously, thus producing a broadening of the peak. The resolution of commercially available instruments of this type is generally limited to about one part in 500. That is, mass 499 can be separated from mass 500 with about a 10 per cent valley. With double-focusing instruments, an electric sector and a magnetic analyzer are placed in tandem to produce both velocity and direction focusing of the ions. Several commercial models of the double-focusing design are available having resolutions in excess of one part in 10000 with an electron impact source, and greater than one part in 3000 with a spark source. The double-focusing geometry is necessary with spark source operation because of the wide energy spread of ions produced in the spark. With both single- and double-focusing, the resolution varies directly with the radius of the analyzer tube and inversely with the width of the slits located in the ion source and ion collector regions. Sensitivity, the abundance of the ions collected per unit sample charge, varies inversely with slit width, and a compromise must be made between resolution and sensitivity. In the T.O.F. design, all ions produced in the source are accelerated through a given electric field and achieve velocities inversely proportional to their masses. Mass separation results from the different times required for various mass ions to traverse the distance to the collector through a linear drift tube. Resolution on the order of 1 part in 300 to 500 can be obtained with this type of mass

analyzer. The major advantage of the T.O.F. instrument is its rapid production and recording of spectra. Over 10000 mass spectra can be produced each second extending through a mass range from 1 to 5000. Other types of mass resolving systems include cycloidal, omegatron, and quadrupole designs.

The two types of ion detection and recording systems are the photographic plate and the electrical detector. The photographic method is commonly used with double-focusing instruments such as the Mattauch-Herzog design which focuses all ions simultaneously in one plane. The photographic plate records a complete spectrum (mass range $\sim 36:1$) in a time interval of a few seconds to 10 minutes. However, the response of the plate to the ion intensity is non-linear and quantitative results are more difficult to obtain than by electrical detection. Electrical detection systems use an ion collector, amplifier, and recorder.

Positive and negative ions and neutral species are produced by the electron bombardment of molecules. The mass spectrum of a compound is a record of the positive ions collected. Positive ions are produced by the removal of one or more electrons from the molecule and by the rupture of one or more bonds, fragmenting the molecule. While the majority of the positive ions are singly charged, doubly and triply charged ions are observed in many instances. Certain mass ions produced from organic molecules must be attributed to the rearrangement of hydrogen atoms during the ionization and fragmentation processes. Metastable ions, formed when ions decompose while traversing the path to the collector, are also frequently observed. Metastable ions generally appear at non-integral mass units.

In the electron impact source, the electron energy is usually adjusted to 50 to 70 eV which is considerably above the appearance potential for molecular and fragment ions. For simplification of a complex spectrum, the bombarding energy can be reduced to provide sufficient energy to ionize the molecule but not enough to rupture bonds, thus achieving a spectrum consisting only of molecular ions. Mass ions appearing in the normal mass spectrum correspond to the various atoms and combinations of atoms in the original molecule. The pattern of mass-ion intensities observed is independent of pressure. Differences in the patterns obtained for various compounds can be used as the basis for the analysis of complex mixtures. Quantitative analysis is based on the ion current varying linearly with the partial pressure of the gas.

Some of the common uses for mass spectrometry include analysis of petroleum products, determination of the structure of organic molecules, determination of trace impurities in gases, residual vacuum studies and leak detection in high-vacuum systems, geological age determinations, tracer techniques with stable isotopes, determination of unstable ionic species in flames, identification of compounds separated by gas-chromatography, trace element analysis in metals

and other solids, microprobe studies of surfaces and studies of surface composition of various materials.

A. G. SHARKEY, JR.

References

1. Duckworth, Henry E., "Mass Spectroscopy," Cambridge, Cambridge University Press, 1958.
2. Beynon, J. H., "Mass Spectrometry and its Applications to Organic Chemistry," Amsterdam, Elsevier Publishing Co., 1960.
3. Biemann, K., "Mass Spectrometry: Organic Chemical Applications," New York, McGraw-Hill Book Co., 1962.
4. McLafferty, F. W., Ed., "Mass Spectrometry of Organic Ions," New York, Academic Press, 1963.
5. McDowell, C. A., Ed., "Mass Spectrometry," New York, McGraw-Hill Book Co., Inc., 1963.
6. Elliott, R. M., Ed., "Advances in Mass Spectrometry," New York, The Macmillan Co., 1963.

Cross-references: IONIZATION, SPECTROSCOPY.

MATHEMATICAL BIOPHYSICS

The term was coined in 1934 by N. Rashevsky to denote a science which applies methods of mathematical physics to biology, just as "biophysics" applies methods of general and experimental physics to biology. Lately a more general term, *mathematical biology*, has been introduced. Mathematical biology stands in the same relation to experimental biology as mathematical physics to experimental physics. This means that mathematical biology develops mathematical theories of various biological phenomena, with the aim of better understanding their nature and of suggesting new avenues for experimental approach. Contrary to mathematical biophysics, however, mathematical biology does not necessarily introduce definite physical models. Some work on mathematical biology may be considered as a construction of purely formal mathematical models of biological phenomena. On the other hand, mathematical biophysics does specifically deal with physical models. Thus mathematical biology includes mathematical biophysics. It is not advantageous to attach specific labels to any particular scientific endeavor, especially in these days of increasing interdisciplinary research. Attaching such labels in science may be just as unwise as attaching definite labels to social and political situations. At best, it may be meaningless. Therefore, in a particular instance it may be difficult, if not impossible, to decide whether a given research belongs to mathematical biophysics or to mathematical biology.

The classical works of Alfred J. Lotka and of Vito Volterra, as well as the similar work by V. A. Kostitzin, on the interaction of species, as well as the wealth of mathematical work on genetics (Sewall Wright, J. B. S. Haldane) should be classed as mathematical biology. In this work no specific physical mechanisms are discussed. It is more formal or phenomenological. On the other

hand, the mathematical work on the models of cardiovascular system, which dates back to Leonard Euler in 1775, and which formed the subject of numerous papers by O. Frank between 1899 and 1928 (*Zeitschrift f. Biologie*), as well as the more recent work of S. Roston and Freeman Cope, J. Womersley, and Allen King on the elasticity of blood vessels, must definitely be considered as mathematical biophysics, even though much of that work was done before the term was coined. Similarly, the work of Braune and Fisher, at the turn of the century, on the dynamics of human locomotion is mathematical biophysics.

Mathematical biology is now frequently referred to as *biomathematics*. The latter term has the advantage of being shorter, but it suffers from a serious etymological shortcoming. One never uses the word *physicomathematics* or *physical mathematics*, instead of mathematical physics. The noun which denotes the principal field of study is *physics*. The word *mathematical* is used, as an adjective, to denote the type of tools used. Mathematics remains the same, whether it is applied in physics, in engineering or in biology. But mathematical physics and mathematical biology are different in the approaches to their problems from experimental physics or experimental biology.

From the point of view of the definition of mathematical biophysics given here, it is not quite as young a science as it appears to be. In fact, biological literature has been for quite some time sporadically sprinkled with studies which are essentially mathematical biophysics.

After coining this word in 1934, N. Rashevsky and his associates at the University of Chicago began a concerted effort of developing mathematical biophysics in a systematic way. For a while this concentrated effort was largely limited to the University of Chicago. Gradually, however, partly due to the migrations of individuals of the group, partly quite independently, work in mathematical biophysics began to spread very widely. The work of the Chicago group became gradually more general, and the many aspects of it are now better described by the more general term *mathematical biology*. There is no entirely comprehensive treatise that encompasses *all* the work done in mathematical biophysics. The references at the end of this article give a limited list of books and journals, some of which deal both with mathematical biophysics as well as more generally with the broader aspects of mathematical biology. The interested reader will find in these references a lead for further study.

In 1939 a special journal was founded by N. Rashevsky, *The Bulletin of Mathematical Biophysics*. It is now in its twenty-seventh year of existence. The *Journal of Theoretical Biology*, founded in 1961 by J. F. Danielli, is devoted to essentially the same subject. A number of important papers on mathematical biophysics has been published in the relatively new *Biophysical Journal* which is devoted to both experimental and theoretical work in biophysics. A great deal of important work in mathematical biophysics is

scattered throughout numerous classical journals in physiology and general biology. Hundreds of papers have been published in the *Bulletin of Mathematical Biophysics* alone. The aim of this article is therefore not to give a detailed mathematical discussion of any particular problems of mathematical biophysics, but rather to mention only *some* of the important work. The work that is not mentioned is omitted only because of lack of space, and not because it is considered as less important.

We have already mentioned the work on cardiovascular phenomena and on the dynamics of human locomotion. The work of Rashevsky began with an unsuccessful attempt at developing a theory of cell division on the basis of so-called diffusion drag forces. Those forces appear in any diffusion field and are due to the interaction between the molecules of the diffusing solute and those of the solvent. Since diffusion phenomena are known to be widespread in cells, the idea was investigated mathematically, as to whether the diffusion drag forces may not be the cause of cell division. Mathematically the problem divided into three parts: the calculation of diffusion fields, the calculation of diffusion drag forces, and the calculation of their possible mechanical effects. It was possible to show that under certain conditions those forces will produce a fission of a cell as a whole. Not only did the theory lead to correct order of magnitude for the average size of cells, but also to a correct quantitative representation of the over-all phenomena of original elongation and eventual construction of such cells as demembrated *Achacia* eggs. Yet the theory failed completely in accounting for the all-important phenomena in the mitotic apparatus. The theory was developed at the time when little was known of the fine phenomena of replication of the smallest parts of the cell. It was based on the physics of "matter in bulk," whereas it is now clear that the whole problem must be attacked on a molecular level.

Nevertheless a number of side problems turned out to be of use. Thus the theoretical study of the limitation of cellular biochemical reactions by the diffusion processes led to a theory of the dependence of the rate of cell respiration on the amount of available oxygen. The theory was found to be in agreement with available data (H. D. Landahl).

An elaboration of the theory of diffusion drag forces led to a theory of self-regulating cell polarity and to a representation of some embryological phenomena.

Another example of the work of the Chicago group is H. D. Landahl's study of the retention of particulate material contained in the air inhaled through the respiratory tracts. This retention is due to a number of physical factors, such as impaction against nasal hairs, impaction against the mucus covered walls of the passages, Brownian movement and sedimentation. The results of this theoretical work are of great practical importance.

One of the most important fields of mathematical biophysics is the theory of biological

membrane potentials, and the theory of transport across biological membranes. The classical work of J. F. Danielli on the structure of membranes must be mentioned here. Theoretical problems of transport across the membrane have been studied by H. H. Ussing, J. Frank and J. E. Meyer, C. Patlak, D. E. Goldman, and many others. The problems deal largely with the phenomenon of the transport of sodium and potassium ions from the side of the membrane where the concentration is lower, to the side where it is higher. Torsten Teorell has studied mathematically transport phenomena in membranes which involve not only the movements of ions but also the movement of water. He has shown, both mathematically and experimentally that periodic fluctuations of the membrane potential do arise under certain conditions. Those potential changes look remarkably similar to potential changes in repetitive nervous discharges, and thus give a possible clue for the understanding of phenomena of nervous excitation.

Kenneth Cole and his associates have contributed important experimental studies of the electrical properties of biological membranes, especially of their impedance. Their work is largely studded with interesting mathematical interpretations, which must be considered as falling into the domain of mathematical biophysics. Interesting physical models, which offer an explanation of the formal equivalence of some biological membranes with what is known as "equivalent electric circuits," have been recently studied by A. Mauro.

Somewhat in a class by itself stands the very important work of A. L. Hodgkin and A. F. Huxley on the mathematical description of nerve excitation. From a series of experiments, they come to the conclusion that the appearance of the action potential is due to the movement of sodium ions inward, into the nerve fiber, which makes the outside negative. The decrease of the action potential is interpreted as due to an outward movement of potassium ions. Empirical relations governing these movements are determined from experiments. The validity of the arguments which lead to those empirical relations is not yet generally recognized. Hodgkin and Huxley then show that from those empirical relations, obtained physiologically under somewhat artificial conditions of a "voltage clamp" which maintains a constant voltage in spite of redistribution of ions, the general shapes of the action potential curves can be calculated. Strictly speaking, the work of Hodgkin and Huxley does not represent a theory in the usual sense of the word. The authors themselves feel that they cannot go back to "first principles." Thus their work, in a sense, represents an empirical mathematical description of some important phenomena by means of a series of very complicated empirical equations. This, in spirit, is rather different from the deductive approach of a mathematical biophysicist. Yet some definite physical pictures are assumed, and their work therefore may be considered as falling at least partly into the domain of mathematical biophysics.

A great deal of mathematical work has been done in biochemistry. It covers, for example, such phenomena as biochemical reaction rates, enzyme activity, etc. To the extent that chemistry itself has now, through quantum mechanics, become a branch of physics, the above-mentioned studies do fall also into the domain of mathematical biophysics. This is enhanced by the fact mentioned in connection with Rashevsky's studies, that purely physical phenomena, such as diffusion, do impose limitations upon, and thus affect, the purely biochemical processes (J. Z. Heaton). This interaction of biophysics and biochemistry is brought to light particularly in the theory of distribution of different metabolites between the different "compartments" of an organism, between which there may be transport of material either by diffusion, convection, or some other mechanisms. This theoretical work is of particular importance for biological studies of movements of metabolites by means of radioactive tracers. In this connection the important work of C. W. Sheppard, J. S. Stevenson, A. Rescigno and G. Segré, and others should be mentioned.

A special application of this type of study has been made to the effects of drugs which are carried to various places in an organism. Here we must mention the beautiful recent work of R. Bellman, R. Kalaba and J. A. Jacquez. More elementary studies of this type date back some 30 years, to W. Gehlen and E. Beccari.

A great deal of mathematical work has been done on the central nervous system. Most of it, however, is of the broader nature of mathematical biology, rather than mathematical biophysics. Nevertheless, the work of W. Rall on the transmission at synapses must be considered as belonging to mathematical biophysics.

The mathematical theory of the regulation of the functions of the lung (J. Defares) should also be mentioned, as well as the work of L. Danziger and G. Elmergreen, and the quite recent work of N. Rashevsky, on oscillatory phenomena in the endocrine system. The size of this article, however, does not permit us to give justice to all the important work that has been done in mathematical biophysics. Instead, we have merely tried to give a general idea of what kind of problems have been studied. The reader will find much more in the references.

N. RASHEVSKY

References

- Rashevsky, N., "Mathematical Biophysics: Physico-mathematical Foundations of Biology," Third edition, 2 volumes, New York, Dover, 1960.
- Rashevsky, N., "Some Medical Aspects of Mathematical Biology," Springfield, Ill., Charles Thomas, 1964.
- Rashevsky, N., Ed., "Physicomathematical Aspects of Biology," in "Proceedings of the International School of Physics 'Enrico Fermi,'" Varenna, Italy, New York and London, Academic Press, 1962.
- Rashevsky, N., Ed., "Mathematical Theories of Biological Phenomena," a symposium, *Ann. N.Y. Acad. Sci.*, **96**, 895-1116 (1962).
- Sheppard, C. W., "Basic Principles of the Tracer Method," New York and London, John Wiley & Sons, 1962.
- Rescigno, A., and Segré, G., "La Cinetica dei Farmaci e dei Traccianti Radioattivi," Torino, Boringhieri, 1961.
- Riggs, D. S., "The Mathematical Approach to Physiological Phenomena," Baltimore, Williams and Wilkins, 1963.
- Grodins, F. S., "Control Theory and Biological Systems," New York, Columbia University Press, 1963.
- Rashevsky, N., Ed., *The Bulletin of Mathematical Biophysics*, published since 1939.
- Danielli, J. F., Ed., *Journal of Theoretical Biology*, published since 1961.
- Oncley, J. L., Ed., *Biophysical Journal*, published since 1960.

Cross-reference: BIOPHYSICS.

MATHEMATICAL PHYSICS

The term "mathematical physics" is almost synonymous with "THEORETICAL PHYSICS," but their difference is significant. It is like the difference between the descriptions of the electromagnetic field by Maxwell and by Faraday respectively. The theoretical (nonmathematical) description draws on analogies between elements of the field and familiar mechanical models—stretched strings, compressed fluids, vortex motion, etc.; the mathematical description made use of the abstract analytical properties of the elements of the field to set up a purely symbolic description without mechanical models. Classical theoretical physics was largely mathematical in content, but was nevertheless based on mechanical models in the spirit of Faraday's theory of the electromagnetic field. The atom and interactions between atoms were regarded as the "real," "external" objects in terms of which all physical phenomena could be explained. The mathematical formalism was merely a handy tool or language in terms of which to set up the explanation. The atoms themselves were not explained, but regarded as the fundamental "building blocks" of the physical world.

Einstein's theory of RELATIVITY is a magnificent historical example of mathematical physics we may cite to contrast with the classical atomic theory. Here mathematical abstractions, Minkowski space 4-vectors, Riemannian tensors, etc. were invented or adopted from the stock-in-trade of pure mathematicians, with analytical properties that were seen to match those of the data of experimental physics—velocities, forces, field variables, etc. Then the logical (i.e., mathematical) consequences of relations among these abstractions predicted new and unexpected relations among either already known or as yet undiscovered data of experimental physics. The construction of a self-consistent mathematical

description of all physical phenomena, without the use of hypothetical building blocks of any kind, is the aim of mathematical physics as distinct from theoretical physics.

The activities of mathematical physicists have resulted in the invention of new mathematical abstractions some of which were at first rejected by pure mathematicians as illogical, only later to be granted a respectable status in the vocabulary of pure mathematics. Examples include Oliver Heaviside's operational calculus, J. Willard Gibbs' vector analysis, and P. A. M. Dirac's delta-function techniques. On the other hand, many branches of pure mathematics which initially had been regarded as so abstract as to be entirely "useless," have been found by mathematical physicists to serve as remarkably useful tools in describing physical phenomena. Examples include non-Euclidean geometry in the problems of COSMOLOGY; function space in modern QUANTUM MECHANICS; spinor analysis, or the theory of binary forms, in quantum FIELD THEORY. Again collaboration between mathematical physicists and mathematicians has in recent years resulted in the construction of new disciplines of great value, examples being group theory, operations analysis, the theory of random functions, information theory, and CYBERNETICS. The names of many contemporary scientists are involved here, including Eugene P. Wigner, John von Neumann, C. E. Shannon, Norbert Wiener and many others.

On closer examination it becomes difficult to distinguish clearly between mathematical physics and applied mathematics; very frequently the same individual may be responsible for discoveries in both areas. Classical examples of this may be cited: Isaac Newton, Laplace, Carl Friedrich Gauss, Henri Poincaré, David Hilbert, Ernst Mach, A. N. Whitehead. Evidently our attempt to define mathematical physics is degenerating into a simple catalog of items with only a vague hint of general characteristics common to all particulars. Physics has sometimes been defined as what physicists do, and one is expected to recognize the physicist without need for further definition than his own affirmation. Mathematical physics may then be defined as what physicists do with mathematics, or what mathematicians do with physics, or some superposition of the two. As the history of mathematical physics unfolds it becomes apparent that activity tends to cluster in a few fruitful directions at any one time. Current interests can be judged from the contents of the leading journals devoted to the subject; among these the reader should consult the *Journal of Mathematical Physics*, and the *Physical Review*, published by the American Institute of Physics; *The Proceedings of the Cambridge Philosophical Society*; *Comptes Rendus* (French Academy of Sciences), *Progress of Theoretical Physics* (Japan), *Nuovo Cimento* (Italy), *Indian Journal of Theoretical Physics*, *Zhurnal Eksperimental'noy i Teoreticheskoy Fiziki* (USSR) (in English Translation "JETP") and other translations published by the American Institute of Physics. Probably the most popular

fields in recent years have been in the wide application to solid-state physics and statistical mechanics of quantum field theoretical techniques introduced initially to deal with the phenomena of high energy physics—nuclear interactions, creation and destruction of particles, etc.

A philosophy of mathematical physics has gradually evolved with all this creative activity. Among the best sources for a study of this the reader is referred to some semi-popular expositions:

Kenneth Ford, "The World of Elementary Particles," Blaisdell Publishing Co., 1963.

James R. Newman, "The World of Mathematics," Simon and Schuster, 1956 (especially Vol. II).

More technical references of a general scope include the classic work in Germany by R. Courant and D. Hilbert, recently translated into English:

Courant and Hilbert, "Methods of Mathematical Physics," Interscience, 1953 and 1962.

Many College courses of theoretical physics are based on the series of volumes by Arnold Sommerfeld, published by Academic Press. A more advanced reference is the two volume set "Methods of Theoretical Physics," by Morse and Feshbach. A less ambitious volume serves as an advanced undergraduate text: "Introduction to Mathematical Physics" by W. Band. An excellent reference work is "The Mathematics of Physics and Chemistry" by Margenau and Murphy in two volumes. Many other references will be found in these two texts.

WILLIAM BAND

MATRICES

Matrix notation and operations are introduced into theoretical physics so that algebraic equations and expressions in terms of rectangular arrays of numbers can be systematically handled.

An $n \times m$ matrix $A = (a_{ij})$ possesses n columns and m rows, having in double suffix notation the form

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

The general element a_{ij} may be a complex number. If all elements are zero, A is the null matrix O or 0 . When $n = 1$, the matrix is a column vector v . If $m = n$, A is square of order n ; if all elements not on the leading diagonal $a_{11}, a_{22}, \dots, a_{nn}$ are zero, the matrix is a diagonal matrix D , while D is the unit matrix I if $a_{11} = a_{22} = \cdots = a_{nn} = 1$.

The sum or difference of two $n \times m$ matrices A and B is an $n \times m$ matrix $C = A \pm B$, where $c_{ij} = a_{ij} \pm b_{ij}$. The elements of αA are αa_{ij} .

The transpose of A is denoted by A' ; this is an $m \times n$ matrix whose i th row and j th column are identical respectively with the j th column and i th row of A . Hence v' is a row matrix with m elements. For convenience, the column v is often printed as a row with braces, $\{v_1 v_2 \cdots v_m\}$.

The product $C = AB$ is only defined when the number of columns of A equals the number of rows of B ; A and B are then conformable for multiplication. If A is $n \times m$ and B is $p \times n$, then C is $p \times m$, with

$$c_{ij} = \sum_{k=1}^n a_{ik}b_{kj}$$

Generally, multiplication is not commutative, but it is always associative. The transpose of a product is given by $(ABC)' = C'B'A'$.

If A is square, then A is symmetric if $A = A'$, while if $A = -A'$ it is skew-symmetric. A quadratic form S_A in the n variables contained in the column $x = \{x_1, x_2, \dots, x_n\}$ may be written as $S_A = x'Ax$, where A is symmetric.

Let $\det A \equiv |A|$ denote the determinant of the square matrix A . If $\det A \neq 0$, A is non-singular. Then the definition of matrix multiplication ensures that

$$\det(AB) = \det A \det B$$

where A and B are square matrices of the same order.

The cofactor of a_{ij} in the square matrix A equals $(-1)^{i+j}$ times the determinant formed by crossing out the i th row and j th column in A . The adjoint of A , denoted by $\text{adj } A$, is the transpose of the matrix formed when each element of A is replaced by its cofactor. We have

$$A \text{ adj } A = (\text{adj } A)A = (\det A)I$$

and $\det(\text{adj } A) = |A|^{n-1}$. The unique reciprocal or inverse of a non-singular matrix A is given by

$$A^{-1} = (\text{adj } A)/\det A$$

This has the property that $AA^{-1} = A^{-1}A = I$. It follows that

$$(AB)^{-1} = B^{-1}A^{-1}$$

Linear equations relating n variables x_i to n variables y_i may be expressed as $x = Ay$; if $\det A \neq 0$, the unique solution for the y_i in terms of the x_i is $y = A^{-1}x$.

The rank of an $n \times m$ matrix A is the order of the largest non-vanishing minor within A ; of the m linear expressions Ax , the rank gives the number that are linearly independent. The m linear equations in n unknowns $Ax = d$, where d is a column with m elements, are consistent if the rank of A equals the rank of the augmented matrix, $(A \ d)$.

If the m linear equations $Ax = d$ are inconsistent, the "best" solution in the least squares sense for x_1, x_2, \dots, x_n is given by the normal equations

$$A'Ax = A'd$$

An $n \times n$ matrix A is orthogonal if $A'A = I$, that is, if $A^{-1} = A'$. Clearly, $|A| = \pm 1$. If c_i denotes the i th column of A , then $c_i'c_i = 1$ and $c_i'c_j = 0$ if $i \neq j$; similar results hold for the rows. The transformation $x_1 = Ax$ represents a rotation of rectangular Cartesian axes in three dimensions;

$|A| = +1$ if the right-handed character is preserved. The element λ_{ij} equals the cosine of the angle between the $(x_1)_i$ and x_j axes. If N is skew-symmetric, then $(I + N)^{-1}(I - N)$ is orthogonal.

Matrices with complex elements are manipulated according to the same rules. A square matrix H is Hermitian if $H' = H$, and skew-Hermitian if $H' = -H$, a star denoting the complex conjugate. A unitary matrix U satisfies $U' = U^{-1}$, the columns (and rows) enjoying the properties $c_i'c_j = 1$ if $i = j$ and 0 if $i \neq j$.

First- and second-order tensors, arising in many physical problems, may be expressed in matrix notation. If $x_1 = Ax$ denotes a rotation of rectangular Cartesian axes, A being orthogonal, then $f_1 = Af$ and $F_1 = AF A'$ define Cartesian tensors of orders 1 and 2 respectively. Evidently, if u and v are vectors or tensors of order 1, then $u'v$ is an invariant, Fu is a tensor of order 1 and uv' is a tensor of order 2. For example, the vector product $u \times v$ may be written as Uv , where

$$U = \begin{pmatrix} 0 & u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}$$

U is a tensor of order 2 if u is a vector; U' is the dual of u .

But if $x_1 = Ax$, where A is not orthogonal, $f_1 = Af$ defines a contravariant vector, but $g_1 = A'g$ defines a covariant vector. The product $g'f$ is now an invariant.

If A is square of order n , then the n homogeneous equations $Ax = \lambda x$ require $\det(A - \lambda I) = 0$ for non-trivial solutions. This characteristic equation possesses n characteristic or latent roots; if they are all distinct, n corresponding characteristic or latent vectors exist. The vector k_i corresponding to the root λ_i may consist of the n cofactors of any row of $A - \lambda_i I$; at least one non-trivial row exists.

The following properties are important. If A is real and symmetric, and if λ_i and λ_j are distinct, then $k_i'k_j = 0$; these two vectors are orthogonal. Again, if A is real and symmetric, the n values of λ are real, but if A is real and skew-symmetric, these n values are pure imaginary. For a real orthogonal matrix A , $|\lambda_i| = 1$ for all i . The characteristic roots of A^{-1} are $1/\lambda_i$, k_i still being the corresponding vectors. If $|\lambda_1|$ is the largest of the moduli of the n roots, then as $r \rightarrow \infty$, $A^r x \rightarrow k_1$, where x is an arbitrary column.

If A is symmetric, n mutually orthogonal characteristic vectors may be found even if the roots are not all distinct. If each vector k_i is normalized, i.e., divided by $\sqrt{(k_i'k_i)}$, then the matrix

$$A = (k_1 \ k_2 \ \dots \ k_n)$$

is orthogonal, and the product $A'A$ equals D , the diagonal matrix consisting of the n roots arranged down its leading diagonal in order. The matrix A is said to be diagonalized.

More generally, if A is a general square matrix of order n , then n independent vectors k_i may be

found corresponding to the n roots if the latter are distinct. Then

$$\mathbf{T} = (\mathbf{k}_1 \ \mathbf{k}_2 \ \dots \ \mathbf{k}_n)$$

transforms \mathbf{A} into diagonal form thus: $\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{D}$. If the n roots are not all distinct, this reduction may or may not be possible; it is always possible if \mathbf{A} is symmetric. Note $\mathbf{A}^n = \mathbf{T}\mathbf{D}^n\mathbf{T}^{-1}$.

Similar remarks apply to Hermitian matrices \mathbf{H} . If $\mathbf{H}\mathbf{k} = \lambda\mathbf{k}$, all the n values of λ are real, and n vectors can always be found such that $\mathbf{k}^*\mathbf{k}_j = \delta_{ij}$. The unitary matrix $\mathbf{U} = (\mathbf{k}_1\mathbf{k}_2 \dots \mathbf{k}_n)$ transforms \mathbf{H} into diagonal form.

Two quadratic forms $S_A = \mathbf{x}'\mathbf{A}\mathbf{x}$, $S_B = \mathbf{x}'\mathbf{B}\mathbf{x}$, where \mathbf{A} and \mathbf{B} are symmetric and of the same order, may be reduced simultaneously to sums of squares. The equations $\mathbf{A}\mathbf{k} = \lambda\mathbf{B}\mathbf{k}$ demand $\det(\mathbf{A} - \lambda\mathbf{B}) = 0$; this possesses n roots λ_i and n corresponding vectors \mathbf{k}_i . If

$$\mathbf{T} = (\mathbf{k}_1 \ \mathbf{k}_2 \ \dots \ \mathbf{k}_n)$$

then $\mathbf{T}'\mathbf{A}\mathbf{T}$ and $\mathbf{T}'\mathbf{B}\mathbf{T}$ are both diagonal. The transformation $\mathbf{x} = \mathbf{T}\mathbf{y}$ yields the two sums of squares $S_A = \mathbf{y}'(\mathbf{T}'\mathbf{A}\mathbf{T})\mathbf{y}$ and $S_B = \mathbf{y}'(\mathbf{T}'\mathbf{B}\mathbf{T})\mathbf{y}$. In particular, if S_A is positive definite, $\mathbf{T}'\mathbf{A}\mathbf{T}$ will equal \mathbf{I} if new columns $\bar{\mathbf{k}}_i$ are used in \mathbf{T} , where $\bar{\mathbf{k}}_i = \mathbf{k}_i/\sqrt{(\mathbf{k}_i'\mathbf{A}\mathbf{k}_i)}$.

Necessary and sufficient conditions for the real quadratic form S_A to be positive definite for all real $\mathbf{x} \neq 0$ are that the n determinants

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \dots, \det \mathbf{A}$$

should be positive. This ensures that the n characteristic roots of \mathbf{A} are all positive.

Finally, matrices may often usefully be partitioned employing matrices within matrices. Multiplication may still be performed provided each individual matrix product is permissible. For example,

$$\begin{pmatrix} \mathbf{a} & \mathbf{b}' \\ \mathbf{c} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{e} & \mathbf{f}' \\ \mathbf{g} & \mathbf{H} \end{pmatrix} = \begin{pmatrix} \mathbf{ae} + \mathbf{b}'\mathbf{g} & \mathbf{af}' + \mathbf{b}'\mathbf{H} \\ \mathbf{ce} + \mathbf{D}\mathbf{g} & \mathbf{cf}' + \mathbf{D}\mathbf{H} \end{pmatrix}$$

where \mathbf{a} , \mathbf{e} are scalars, \mathbf{b} , \mathbf{c} , \mathbf{f} , \mathbf{g} are 1×3 columns and \mathbf{D} , \mathbf{H} are 3×3 .

Applications. Differential Equations. If $d\mathbf{x}/dt + \mathbf{A}\mathbf{x} = \mathbf{f}$, \mathbf{A} being constant and $\mathbf{f} = \{f_1(t), f_2(t), \dots, f_n(t)\}$, then if \mathbf{T} diagonalizes \mathbf{A} , $\mathbf{x} = \mathbf{T}\mathbf{y}$ yields n non-simultaneous equations $dy_i/dt + \lambda_i y_i = \mathbf{T}^{-1}\mathbf{f}$. If $y_0(t)$ is a particular integral,

$$\mathbf{x}(t) = \mathbf{T} \begin{pmatrix} e^{-\lambda_1 t} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e^{-\lambda_n t} \end{pmatrix} \times [\mathbf{T}^{-1}\mathbf{x}(0) - \mathbf{y}_0(0)] + \mathbf{T}\mathbf{y}_0(t)$$

Geometry. In three-dimensional Cartesian co-ordinates,

$$\mathbf{a}'\mathbf{x} + d = 0$$

represents a plane, the perpendicular distance from \mathbf{x}_1 being

$$(\mathbf{a}'\mathbf{x}_1 + d)/\sqrt{(\mathbf{a}'\mathbf{a})}$$

The equation $\mathbf{x}'\mathbf{A}\mathbf{x} = d$ represents a central quadric. If \mathbf{A} diagonalizes \mathbf{A} , the rotation $\mathbf{x} = \mathbf{A}\mathbf{x}_1$ yields $\mathbf{x}_1'\mathbf{D}\mathbf{x}_1 = d$. The vectors \mathbf{k}_1 , \mathbf{k}_2 , \mathbf{k}_3 specify the three principal axes, of semi-lengths $\sqrt{(d/\lambda_i)}$ when $d/\lambda_i > 0$.

Dynamics. The rotational equations of motion of a rigid body with respect to moving axes fixed in the body and with the origin fixed in space or at centre of mass are

$$\mathbf{g} = \mathbf{J}\dot{\boldsymbol{\omega}} + \boldsymbol{\Omega}\mathbf{J}\boldsymbol{\omega}$$

where \mathbf{g} = couple, $\boldsymbol{\omega}$ = angular velocity, $\boldsymbol{\Omega}' = \text{dual } \boldsymbol{\omega}$, \mathbf{J} denotes the inertia tensor $= \sum m\mathbf{X}\mathbf{X}$, where $\mathbf{X}' = \text{dual } \mathbf{x}$. Explicitly,

$$\mathbf{J} = \begin{pmatrix} A & H & G \\ H & B & F \\ G & F & C \end{pmatrix}$$

The rotational kinetic energy is $\frac{1}{2}\boldsymbol{\omega}'\mathbf{J}\boldsymbol{\omega}$. When principal axes of inertia are chosen, \mathbf{J} is diagonal, yielding Euler's equations.

Small oscillations about a position of equilibrium are investigated by considering the second order approximations

$$K.E. = \dot{\mathbf{q}}'\mathbf{A}\dot{\mathbf{q}}, \quad P.E. = \mathbf{q}'\mathbf{B}\mathbf{q}$$

\mathbf{q} containing n generalized coordinates measured from their equilibrium values. \mathbf{A} and \mathbf{B} are constant symmetric matrices. If the n roots of $\det(\mathbf{A} + \lambda\mathbf{B}) = 0$ are considered, and if $\mathbf{q} = \mathbf{T}\mathbf{x}$, where $\mathbf{T} = (\mathbf{k}_1\mathbf{k}_2 \dots \mathbf{k}_n)$ reduces \mathbf{A} to the unit matrix \mathbf{I} , the equations of motion are

$$\ddot{\mathbf{x}}_i + (1/\lambda_i)\mathbf{x}_i = 0$$

The elements of \mathbf{x} are the *normal coordinates*; each individual solution x_i in terms of the q 's is a *normal mode* of period $2\pi\sqrt{\lambda_i}$.

Electromagnetic Theory. Maxwell's 3×3 stress tensor in matrix notation is

$$\mathbf{T} = \frac{1}{2}[2\epsilon\mathbf{e}\mathbf{e}' + 2\mu\mathbf{h}\mathbf{h}' - \epsilon(\mathbf{e}'\mathbf{e})\mathbf{I} - \mu(\mathbf{h}'\mathbf{h})\mathbf{I}]$$

in mks rationalized units. The field exerts a force across an area element $\mathbf{n} \delta S$ equal to $\mathbf{T}\mathbf{n} \delta S$.

When electromagnetic waves are propagated in an ionized medium the equation

$$\text{curl curl } \mathbf{e} = k^2(\mathbf{I} - \mathbf{M})\mathbf{e}$$

arises, where

$$\mathbf{M} = X(Y^2\mathbf{nn}' + iY\mathbf{N} - \mathbf{I})/(Y^2 - 1)$$

in the usual notation with collisions neglected; here, \mathbf{n} = unit vector directed along the external magnetic field, $\mathbf{N}' = \text{dual } \mathbf{n}$. These equations may be rearranged in terms of the matrix

$$\mathbf{f} = \{E_x, E_y, Z_0H_x, Z_0H_y\}$$

giving $d\mathbf{f}/dz = -ik\mathbf{T}\mathbf{f}$, where \mathbf{T} is a 4×4 matrix. If the characteristic roots $\lambda_i(z)$ of \mathbf{T} are found, and if \mathbf{R} diagonalizes \mathbf{T} , then the transformation $\mathbf{f} = \mathbf{R}\mathbf{g}$ yields

$$\frac{d\mathbf{g}}{dz} = -ik\mathbf{D}\mathbf{g} \quad \mathbf{R}^{-1} \frac{d\mathbf{B}}{dz} \mathbf{g}$$

The solutions of the approximate equations $d\mathbf{g}/dz = -ikD\mathbf{g}$ are related to the characteristic waves propagated in the medium.

Special Relativity. If $\mathbf{x} = \{ict, x, y, z\}$ refers to an inertial frame S , and if a second parallel frame S_1 has uniform relative velocity U along Ox , the Lorentz transformation is $\mathbf{x}_1 = \Lambda_1 \mathbf{x}$, where

$$\Lambda_U = \begin{pmatrix} \beta & iU\beta/c & 0 & 0 \\ iU\beta/c & \beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Λ is orthogonal, and $\beta = 1/\sqrt{1 - U^2/c^2}$. We have $\Lambda_1 \Lambda_{1'} = \Lambda_{11'}$, where

$$W = (U + V)/(1 + UV/c^2).$$

For the general velocity \mathbf{v} relating parallel frames,

$$\Lambda = \begin{pmatrix} \beta & i\beta\mathbf{v}/c \\ i\beta\mathbf{v}/c & 1 + (\beta - 1)\mathbf{v}\mathbf{v}'/c^2 \end{pmatrix}$$

The operator $\square = \{\partial/\partial ict, \partial/\partial x, \partial/\partial y, \partial/\partial z\}$ is a four vector satisfying $\square_1 = \Lambda \square$; so are the four-current \mathbf{i} and the four-potential \mathbf{b} ,

$$\mathbf{i} = \begin{pmatrix} ic\rho \\ \mathbf{j} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} i\phi/c \\ \mathbf{a} \end{pmatrix}$$

where \mathbf{a} is the vector potential. They satisfy $\square \cdot \mathbf{i} = 0$ (conservation of charge), $\square \cdot \mathbf{b} = 0$ (the Lorentz relation). Maxwell's equations in mks units in free space take the form

$$\square \cdot \mathbf{F} = i/\epsilon_0 c^2$$

$$\square \cdot \mathbf{G} = 0$$

where

$$\mathbf{F} = \square \mathbf{b}' - (\square \mathbf{b})' = \begin{pmatrix} 0 & ie'/c \\ ie'/c & \mu_0 \mathbf{h} \end{pmatrix}$$

and

$$\mathbf{G} = \begin{pmatrix} 0 & \mu_0 \mathbf{h}' \\ -\mu_0 \mathbf{h} & i\mathbf{E}/c \end{pmatrix}$$

are tensors of order 2 under a Lorentz transformation. \mathbf{E}' and \mathbf{h}' are the respective 3 \times 3 duals of \mathbf{e} and \mathbf{h} . All tensor equations are invariant in form in all frames of reference.

The tensor

$$\mathbf{T} = \frac{1}{2}\epsilon_0 c^2 \mathbf{F} \mathbf{F} - \frac{1}{2}\mu_0 {}^{-1}\mathbf{G} \mathbf{G}$$

$$= \frac{1}{2} \begin{pmatrix} \epsilon_0 \mathbf{e}' \cdot \mathbf{e} + \mu_0 \mathbf{h}' \cdot \mathbf{h} & 2ie'\mathbf{h}/c \\ 2i\mathbf{E}\mathbf{h}/c & \epsilon_0 \mathbf{e}\mathbf{e}' + \epsilon_0 \mathbf{E}\mathbf{E}' + \mu_0 \mathbf{h}\mathbf{h}' + \mu_0 \mathbf{H}\mathbf{H}' \end{pmatrix}$$

contains the energy density, Poynting's vector, the momentum density and Maxwell's stress tensor in partitioned form.

Applications may likewise be made to circuit theory, to elasticity where 3×3 stress and strain tensors are defined, and to quantum mechanics, embracing, for example, matrix mechanics and the Dirac wave equation of the electron.

JOHN HEADING

References

- Heading, J., "Matrix Theory for Physicists," London, Longmans, Green & Co.
 Heading, J., "Electromagnetic Theory and Special Relativity," Cambridge, University Tutorial Press.
 Jeffreys, H., and Jeffreys, B., "Methods of Mathematical Physics," Cambridge, The University Press.
 Perlis, S., "Theory of Matrices," Reading, Mass., Addison-Wesley.

Cross-references: MATRIX MECHANICS.

MATRIX MECHANICS

Introduction. This article is a sequel to the article WAVE MECHANICS which appears in this Encyclopedia. We assume that the reader is familiar with the content of that article, which will be referred to here as WM.

The point of view adopted in WM is that non-relativistic quantum mechanics can be formulated entirely in terms of the time-dependent Schrödinger wave equation. While this is true, there are, as remarked in the introduction of WM, situations in which the matrix approach, or a combination of matrix and wave methods, is useful. In the present article, we first discuss some properties of matrices and show how they appear in quantum mechanics. Purely matrix methods are then used to calculate the energy levels and some other properties of the linear harmonic oscillator (WM, p. 773).

Matrices in Quantum Mechanics. A matrix is a square or rectangular array of numbers. Two or more matrices may be added or multiplied in accordance with certain rules. As a simple example, consider the matrix A , which has three rows and four columns; we write it:

$$A = \begin{pmatrix} (1|A|1) & (1|A|2) & (1|A|3) & (1|A|4) \\ (2|A|1) & (2|A|2) & (2|A|3) & (2|A|4) \\ (3|A|1) & (3|A|2) & (3|A|3) & (3|A|4) \end{pmatrix} \quad (1)$$

The twelve parenthetical symbols on the right side of Eq. (1) are the *matrix elements*, and in general are complex numbers which may be functions of some variable such as the time. Two matrices can be added only if they have the same number of rows and the same number of columns; addition then consists in forming the matrix whose elements are the sums of the corresponding elements of the original matrices:

$$C = A + B \text{ implies } (n|C|m) = (n|A|m) + (n|B|m) \quad (2)$$

Two matrices can be multiplied only if the number of columns of the left member of the product is equal to the number of rows of the right member:

$$C = AB \text{ implies } (n|C|m) = \sum_k (n|A|k)(k|B|m) \quad (3)$$

It is apparent from Eqs. (2) and (3) that addition is commutative, but multiplication in general is not.

A *unit matrix* is one which can multiply another matrix and leave it unchanged. It follows from Eq. (3) that it is a square matrix (equal number of rows and columns) that has unity along its principal diagonal (upper left to lower right) and zero elsewhere:

$$(n|1|m) = \delta_{nm} = 1 \text{ if } n = m \text{ and } 0 \text{ if } n \neq m \quad (4)$$

δ_{nm} is the Kronecker δ -symbol. A *constant matrix* $C = c1$ has the number c along the principal diagonal and zero elsewhere. A *diagonal matrix* D has the form:

$$(n|D|m) = D_n \delta_{nm} \quad (5)$$

The numbers D_n are called the eigenvalues of the matrix D .

A matrix A may or may not possess an inverse A^{-1} , which must satisfy both of the relations:

$$AA^{-1} = 1, \quad A^{-1}A = 1$$

The Hermitian adjoint A^\dagger of a matrix A is the matrix formed by interchanging rows and columns and taking the complex conjugate of each element:

$$B = A^\dagger \text{ implies } (n|B|m) = (m|A|n)^* \quad (6)$$

A matrix is *Hermitian* or self-adjoint if it is equal to its Hermitian adjoint: $A = A^\dagger$. It follows from Eq. (6) that only square matrices can be Hermitian. A matrix is *unitary* if its Hermitian adjoint is equal to its inverse:

$$A^\dagger = A^{-1}, \text{ or } AA^\dagger = 1 \text{ and } A^\dagger A = 1 \quad (7)$$

The transformation of a square matrix A into a square matrix A' by means of the transformation matrix S is defined by:

$$SAS^{-1} = A' \quad (8)$$

It is evident that S^{-1} must exist, and that it transforms A' back into A . The form of a matrix equation is unaffected by transformation. Thus the equation

$$AB = CDE = F$$

may be transformed into

$$SABS^{-1} = SC'DE'S^{-1} = SFS^{-1}$$

which is equivalent to

$$A'B' = C'D'E' = F'$$

where the primed and unprimed matrices are related by Eq. (8). This invariance of matrix equations with respect to transformations makes it possible to work with any convenient transformation of a set of matrices without affecting the validity of any results obtained.

It can be shown* that any Hermitian matrix

* The phrase "it can be shown," which appears occasionally in this article, means that the necessary proofs are too long to be given here. They may be found in the books listed as references at the end of the article.

can be transformed into diagonal form by means of a suitable unitary transformation matrix. The diagonal elements of the resulting diagonal matrix are called the eigenvalues of the original Hermitian matrix, as well as of the transformed matrix [see Eq. (5)]. These eigenvalues are easily seen to be real numbers. Since the eigenvalues are the measurable values of quantities represented by matrices, we shall require that physical quantities such as x , p , and H , which must have real eigenvalues, correspond to Hermitian matrices. It can also be shown that the necessary and sufficient condition that two Hermitian matrices A and B can be diagonalized by the same unitary transformation is that they commute, i.e., that $AB = BA$.

Suppose now that the Schrödinger equation [WM, Eq. (14)] is written in the form:

$$H\psi_m = E_m\psi_m \quad (9)$$

where E_m are the energy levels or eigenvalues and the ψ_m are the corresponding *eigenfunctions*. The eigenfunctions can not only be chosen to be normalized [WM, Eq. (19)], so that $\int \psi_m^* \psi_m d\tau = 1$, but also orthogonal to each other: $\int \psi_n^* \psi_m d\tau = 0$ if $n \neq m$. Thus if Eq. (9) is multiplied through by ψ_n^* and integrated over all coordinates, we obtain:

$$\int \psi_n^* H \psi_m d\tau = E_m \delta_{nm} \quad (10)$$

We define the matrix element of the Hamiltonian H in the representation specified by the functions ψ_m as:

$$(n|H|m) = \int \psi_n^* H \psi_m d\tau \quad (11)$$

Equation (10) then shows that when the eigenfunctions ψ_m are used to define the representation, the matrix of H is diagonal, and the diagonal elements are the energy levels of the system.

If some other set of functions ϕ_k , also normalized and orthogonal so that $\int \phi_k^* \phi_l d\tau = \delta_{kl}$, is chosen to specify the representation, the matrix for H would not be in diagonal form. That is, the matrix element

$$(k|H|l) = \int \phi_k^* H \phi_l d\tau \quad (12)$$

would not vanish when $k \neq l$. Now it can be shown that the transformation matrix S , defined by

$$(n|S|k) = \int \psi_n^* \phi_k d\tau \quad (13)$$

is unitary, and transforms the nondiagonal matrix of Eq. (12) into the diagonal matrix of Eq. (11). That is, the transformation equation [Eq. (8)], with the substitution of Eq. (7), may be written:

$$SHS^\dagger = H'$$

or in matrix element form:

$$\sum_k \sum_l (n|S|k)(k|H|l)(l|S^\dagger|m) = (n|H|m) = E_m \delta_{nm}$$

We see then that there are two general methods for finding the energy levels of a system. One is to solve the Schrödinger equation, as in WM, and

determine the energy eigenvalues. The other is to choose a convenient set of functions ϕ_k , use them to calculate a matrix representation [Eq. (12)] for the Hamiltonian, and transform the resulting matrix into diagonal form. The eigenvalues of the diagonal form of H are then the energy levels. It may happen that H can be put in diagonal form directly, without explicit reference to a set of functions ϕ_k or the transformation matrix [Eq. (13)]. An example of this is given in the next section.

Linear Harmonic Oscillator. The linear or one-dimensional harmonic oscillator is described by the Hamiltonian [WM, Eq. (21)]:

$$H = (1/2m)(p^2 + m^2\omega^2 x^2) \quad (14)$$

where $\omega/2\pi$ is the classical frequency of the oscillator. We also know that [WM, Eq. (15)]:

$$px - xp = \hbar/i \quad (15)$$

We then see from the algebra of Eqs. (14) and (15) that

$$(p + im\omega x)(p - im\omega x) = p^2 + m^2\omega^2 x^2 + im\omega(xp - px) \\ = 2mH - m\hbar\omega \quad (16)$$

and that

$$(p - im\omega x)(p + im\omega x) = 2mH + m\hbar\omega \quad (17)$$

If now we multiply Eq. (16) through on the right and Eq. (17) through on the left by $(p + im\omega x)$, we obtain

$$(2mH - m\hbar\omega)(p + im\omega x) = (p + im\omega x) \\ (p - im\omega x)(p + im\omega x) \quad (18) \\ = (p + im\omega x)(2mH + m\hbar\omega)$$

We choose a representation in which H is diagonal, so that, from Eqs. (10) and (11),*

$$(n|H|n') = E_n \delta_{nn'}$$

In this representation, a non-operator number, such as the right side of Eq. (15) or the term $m\hbar\omega$ that is added to or subtracted from $2mH$ in Eq. (18), must be thought of as a constant matrix [defined under Eq. (4)]. Thus if we write the left side of Eq. (18) out in terms of matrix elements, we encounter the term

$$(n|2mH - m\hbar\omega|n') = (n|2mH|n') - (n|m\hbar\omega 1|n') \\ = (2mE_n - m\hbar\omega)\delta_{nn'}$$

It follows that the left side of Eq. (18) is

$$(2mE_n - m\hbar\omega)(n|p + im\omega x|n')$$

and the right side is

$$(n|p + im\omega x|n')(2mE_{n'} + m\hbar\omega)$$

* We use the matrix index n' instead of m , in order to avoid confusion with the oscillator mass.

On equating these two expressions, rearranging terms, and cancelling a common factor $2m$, we obtain:

$$(E_n - E_{n'} - \hbar\omega)(n|p + im\omega x|n') = 0 \quad (19)$$

Equation (19) tells us that either $(E_n - E_{n'} - \hbar\omega)$ is zero, or $(n|p + im\omega x|n')$ is zero, or both. To make use of this result, we write the n th diagonal matrix element of Eq. (16):

$$\sum_n (n|p + im\omega x|n')(n'|p - im\omega x|n) \\ = 2m(E_n - \frac{1}{2}\hbar\omega) \quad (20)$$

Now p and x are physical quantities, and so correspond to Hermitian matrices. Thus $p = p^\dagger$, $x = x^\dagger$, and $(p - im\omega x) = (p + im\omega x)^\dagger$, so that from Eq. (6):

$$(n'|p - im\omega x|n) = (n|p + im\omega x|n')^* \quad (21)$$

Substitution of Eq. (21) on the left side of Eq. (20) shows that each term in the sum is non-negative, so that the right side is also non-negative. From Eq. (19), not more than one term in the sum can fail to vanish; this is the term for which $E_{n'} = E_n - \hbar\omega$, if there is such a term. If there is no such term, or in other words if E_n is the lowest energy eigenvalue, then the left side of Eq. (20) is zero, and $E_n = \frac{1}{2}\hbar\omega$. If there is such a term, then $E_n = \frac{1}{2}\hbar\omega$, and $E_n - \hbar\omega$ is another eigenvalue.

We conclude that the energy levels of the linear harmonic oscillator form the infinite sequence:

$$\frac{1}{2}\hbar\omega, \frac{3}{2}\hbar\omega, \frac{5}{2}\hbar\omega, \dots$$

or:

$$E_n = (n + \frac{1}{2})\hbar\omega, \quad n = 0, 1, 2, \dots \quad (22)$$

This is in agreement with the lowest energy level obtained in WM, Eq. (23).

The diagonal matrix elements of a physical quantity are the *expectation values* of that quantity [see WM, Eq. (20)]. From the preceding discussion, we see that

$$(n|p + im\omega x|n) = 0 \text{ and } (n|p - im\omega x|n) = 0$$

so that, by addition and subtraction of these equations,

$$(n|p|n) = 0 \text{ and } (n|x|n) = 0$$

Thus the expectation value of x and of $p = (\hbar/i)(d/dx)$ is zero for each oscillator eigenfunction. This is easily verified explicitly for the ground state wave function [WM, Eq. (22)]. However, the expectation values of p^2 and of x^2 are not zero, as we now show.

We know from Eq. (19) that $(n - 1|p + im\omega x|n) = 0$, and hence from Eq. (21) that

$$(n|p - im\omega x|n - 1) = 0 \quad (23)$$

We also know from the derivation of Eq. (22) that $(n|p + im\omega x|n - 1)$ is not in general zero, and we set it equal to $2A_n$:

$$(n|p + im\omega x|n - 1) = 2A_n \quad (24)$$

Adding and subtracting Eqs. (23) and (24) we obtain

$$(n|p|n-1) = A_n \quad \text{and} \quad (n|x|n-1) = -iA_n/m\omega$$

Since p and x are Hermitian, we have from Eq. (6) that

$$(n-1|p|n) = A_n^* \quad \text{or} \quad (n|p|n+1) = A_{n+1}^*,$$

and that

$$(n-1|x|n) = iA_n^*/m\omega \quad \text{or} \quad (n|x|n+1) = iA_{n+1}^*/m\omega$$

From Eq. (3):

$$(n|p^2|n) = \sum_{n'} (n|p|n')(n'|p|n) \\ = (n|p|n-1)(n-1|p|n) + (n|p|n+1)(n+1|p|n)$$

since all other matrix elements in the sum are zero. Thus

$$(n|p^2|n) = |A_n|^2 + |A_{n+1}|^2 \quad (25)$$

and in similar fashion

$$(n|x^2|n) = (|A_n|^2 + |A_{n+1}|^2)/m^2\omega^2 \quad (26)$$

Substitution of Eq. (24) into Eq. (20), with the help of Eq. (22), gives:

$$4|A_n|^2 - 2m(E_n - \frac{1}{2}\hbar\omega) = 2nm\hbar\omega$$

so that Eqs. (25) and (26) yield:

$$(n|p^2|n) = (n + \frac{1}{2})m\hbar\omega, \quad (n|x^2|n) = (n + \frac{1}{2})(\hbar/m\omega) \quad (27)$$

Comparison with Eq. (14) shows that the expectation values of the kinetic and potential parts of the energy are equal, and that each is half the energy eigenvalue. The results of Eq. (27) for the lowest state ($n=0$) can be verified by using Eqs. (20) and (22) of WM, normalizing ψ_0 , and performing the integrations.

Conclusion. These two articles, WM and the present one, show in some detail how the Schrödinger wave equation and the matrix theory are related in a particular, relatively simple case. In more complicated situations it may be most convenient to use one or the other of these approaches, or, most often, a combination of the two. The reference books listed below give many examples of the applications of the theory described thus far, and also of its extension to relativistic motion.

LEONARD I. SCHIFF

References

- Dirac, P. A. M., "The Principles of Quantum Mechanics," Fourth edition, London, Oxford Press, 1958.
Schiff, L. I., "Quantum Mechanics," Second edition, New York, McGraw-Hill Book Co., 1955.

Bohm, D., "Quantum Theory," Englewood Cliffs, N.J., Prentice-Hall, 1951.

Landau, L., and Lifshitz, E., "Quantum Mechanics, Non-Relativistic Theory," Reading, Mass., Addison-Wesley, 1958.

Messiah, A., "Quantum Mechanics," New York, Interscience Publishers, 1961.

Cross-references: MATRICES, QUANTUM THEORY, SCHRÖDINGER EQUATION, WAVE MECHANICS.

MEASUREMENTS, PRINCIPLES OF

Physics in its most fundamental aspect is an experimental science in the sense that speculation is guided by experience; its development depends on frequent comparison between observation and what is deduced from a hypothesis, itself formed to coordinate previous observations. A scientific experiment is a set of events deliberately arranged to reveal as clearly as possible, some regularity in the behavior of physical objects. Compared with other sciences, fundamental physics deals with simple systems, ranging from elementary particles to relatively simple arrangements of atoms. Its experiments are in principle simpler than those of biological science, where the experimenter interferes at his peril with the complicated systems which he studies, or those of cosmology, where the astronomer cannot interfere experimentally and is confined to pure observation of his enormous and remote fields of study. In fact, of course, the theoretical methods of the physicist are sophisticated and the experimental methods are often more elaborate than those employed in other sciences. This elaboration arises from two features of contemporary experimentation, the need to produce extraordinary circumstances, for example, streams of particles having unusually high velocities, and the need to isolate the phenomenon under investigation from all interference. For example, some of the phenomena of the solid state can only be profitably studied in crystalline matter of extreme purity: in many experiments in physics a good vacuum allows particles to travel a long distance without collision or a better vacuum allows a prepared surface to remain for an appreciable time free from contamination by foreign atoms.

Besides fundamental physics, there is a very large field of applied physics, passing imperceptibly into fundamental physics and into engineering; in this field, the systems studied are often man-made devices, and the object of experiments is to obtain data required for design purposes and to examine the performance of finished devices with a view to employing or improving them.

In both fundamental and applied fields, the results of an experiment usually take the form of a set of numerical data which are to be compared with a theoretical calculation, or with another set of experimental data. The value of an experiment may often be judged by the trustworthiness of a numerical result, which in turn depends on the probability of a certain difference (error)

between the result and the result of an ideal experiment in which the required quantity is measured to the exclusion of all disturbance.

The easiest kind of experimental error to estimate and to eliminate is the kind called random; this may be caused by external disturbances or by inherent fluctuations of the kind we shall call noise. It is distinguished by the property that if the experiment is repeated, a different result is obtained, and if it is repeated many times, the difference between the means of groups of repetitions decreases with the number of results in each group. It is clear that the reduction of such errors depends essentially on averaging. There are well-defined experimental and mathematical methods for estimating these errors and reducing them in the most economical way. The accumulation of data for averaging is time-consuming, and since the averaging process necessarily smooths out short-term phenomena, it attenuates the high-frequency sensitivity of the experiment.

A much more serious difficulty arises in determining how far the conditions of an experiment agree with those assumed theoretically to exist. The distinctive skill of an experimental physicist lies in eliminating by design or calculation the systematic error which arises when some unwanted external or internal factor interacts in a consistent way with the system which is being studied. Effects of this kind are not necessarily detected or eliminated by simple repetition of the experiment, and there are no comprehensive rules for designing an experiment free from systematic error. It is sometimes possible to conduct an experiment in such a way that a potentially systematic error appears as a random error which can more easily be detected and eliminated. For example, in photographic photometry, a number of spectra are recorded on a photographic plate, and if the spatial order of the records on the plate coincides with the regular change of some variable in the experiment, a regular variation of plate sensitivity will give rise to an apparently systematic error, which could be avoided by arranging the spectra in random order. Effective techniques for randomizing have been worked out for more complicated situations than this and especially in connection with field trials in agriculture and product control in industry. They are available in books on the design of experiments. As a further application of this principle, if the characteristics of an experiment are liable to change with time, it is undesirable to change a variable of the experiment in a monotonic way since the drift then appears as an undetectable systematic error. It would have been apparent if variations had been made in random order. Apart from this possibility, external disturbances may sometimes be eliminated by means of isolation as discussed below; external conditions may be varied over as wide a range as practicable in order to detect unsuspected interactions, and conditions may then be held as nearly constant as possible during the important observations. Of course, the experiment must be carefully and repeatedly reviewed theoretically to search for possible disturbing

effects, and subsidiary experiments or deliberate variations of the main experiment may be required to detect them. In the final resort, it may be possible to invent a radically different experiment to measure a required quantity or substantiate a theoretical conclusion, the hope being that quite different systematic errors will then be encountered. It is true that most of the fundamental ideas of physics and most of the important numerical quantities used in it rest upon broad experimental foundations rather than single experiments. From time to time, single experiments assume a crucial status, decisive between two theories, as the Bothe-Geiger coincidence experiment of the thirties decided between the electromagnetic and the charged-particle views of penetrating cosmic rays, or the experiments of Wu, Ambler and others, settled the non-conservation of parity. However, such experiments have always appeared, in retrospect at least, as fitting into the broader background of experimental fact.

Every event in an experiment is in principle coupled with events in the outside world, and this coupling may lead to systematic or to random error. The coupling may be mechanical, thermal, or electromagnetic, or it may involve the incursion of radiation or of nuclear particles. Time variations of these disturbances are more important than the steady components, which can be eliminated by proper measurement techniques. The time variations are sometimes produced by the spontaneous variations (fundamental noise) associated with the disturbance. For example, the effective sensitivity of thermal detectors of radiation may be limited by the fluctuations in the (parasitic) exchange of radiation between detector and its immediate surroundings, and the fluctuations can be reduced by operating at a low temperature. Not all time variations are of this fundamental type, and it may be necessary to stabilize the environment (e.g., temperature, pressure, electrical supplies) by the established feedback techniques.

We now review some of the techniques for isolating an experiment from the disturbances considered above.

(a) *Attenuation of mechanical vibrations:* traditionally this is achieved by special building construction, placing mechanically critical experiments on ground or subterranean floors, providing special foundations not closely coupled to the rest of the building; these measures are applicable to astronomical telescopes, but their importance in the physical laboratory is probably decreasing. Mechanically vulnerable assemblies can be isolated by systems of masses, springs and dampers designed as mechanical filters, but there are practical difficulties in securing high attenuation at low frequencies. Sometimes attenuation is required in a particular direction, and vibrations in other directions are relatively unimportant; the design of an antivibration support may then be simplified, as where an antivibration mounting is provided for a rotational instrument like a galvanometer. Finally, rotational disturbances of a

suspended system can be produced by sideways accelerations if the rotating system does not satisfy the conditions for dynamic balance, and these conditions must be satisfied to improve steadiness.

(b) *Thermal effects* on measuring equipment may involve the actual temperature of a particular element, in which case thermostatic devices are required, and one of the most precise is the use of a phase-equilibrium bath, i.e., ice or some other solid in equilibrium with its liquid. But thermal effects may be produced by the appearance of temperature *differences* producing mechanical deformation, thermoelectric voltages, or changes in relative values of resistances. The reduction of these differences demands good thermal transfer among the parts themselves and poor thermal transfer between the parts and the external disturbances, so that insulation must be combined with good conducting layers, or even more effective, stirred liquid baths.

(c) *Magnetic systems* are protected from disturbances in two different ways: by shielding and by astatic design. The former involves layers of soft magnetic material, but it must be remembered that the materials of highest permeability (e.g., the permalloys) saturate in quite weak fields and then produce little or no attenuation of disturbances. They must generally be used within outer shields of substances like soft iron. The second method depends on constructing the measuring system (which may consist of suspended magnets or wound coils) in such a way that the effects of a disturbing field, uniform in space, neutralize one another while the effects to be measured do not. It is clear that the sensitivity to the gradient of a disturbing field depends on detailed design, e.g., the nearness of two coils, and that more complicated systems could be used to remove derivatives of higher order.

(d) *Electrostatic and electromagnetic disturbances* are often easy to eliminate by metallic shielding, but the effect of very small holes in the shielding may be appreciable. The effects are worst in electronic devices of high impedance, particularly electron tubes used as electrometers, which can rectify and detect high-frequency disturbances. Such sensitive equipment may require metallic shielding of nearly gastight quality or, indeed, complete double shielding. Furthermore if the sensitive circuit forms a loop coupled with a varying magnetic flux, disturbances will be induced in it in spite of metal shielding of finite conductivity, and such loops must be avoided.

When the experiment has been protected from disturbances, the next step in measurement technique is to eliminate the steady or slowly varying effects which are still present; the simplest example is that of taking the zero of an instrument before and after a reading. If these readings are taken in a regular time sequence, a linear drift can be completely eliminated. It is important to arrange the apparatus so that "on" and "off" readings, more generally "condition A" and "condition B," may be alternated *rapidly* and

with *minimal disturbance* of conditions. This is often a distinctive feature of a well-designed measurement technique. A fundamentally similar but more sophisticated procedure is obtained by automatic alternation of conditions A and B, followed by automatic separation and subtraction of the results of the two conditions. The simplest example is the use of a "chopper" amplifier to measure quasi-steady voltages; a vibrating switch or its equivalent interrupts the input to an amplifier which is required only to amplify the alternating component and which can be free from long-period drift. The output of the amplifier is measured by a rectifying instrument or, more advantageously, by an instrument which is sensitive only to components properly related in phase to the alternations of the input.

Developments of this technique are far reaching, and there are two important extensions of it. The output obtained can be used to control condition A or condition B until they are equal; this is an application of the feedback principle. Furthermore, since the final output is a zero-frequency measure of the difference between A and B, it can be integrated over a long time by electrical, electronic or electromechanical methods (condenser charging, feedback integrator or integrating motor), and noise-type fluctuations, alternating disturbances outside an extremely small frequency band, and disturbances common to A and B—can be eliminated to a spectacular degree. The accuracy of every measurement is limited by fundamental or fluctuation phenomena which it is convenient to call *noise*, though often it is not practically necessary or technically possible to take the elimination of other disturbances to the point where this noise becomes limiting. If we are concerned with the measurement of a quasi-steady quantity, noise can always be removed by time-averaging, though this process will attenuate small, nonrecurring transitory events, and these must remain permanently masked by the noise.

The most fundamental limitations of this kind arise from the "uncertainty" limitations of quantum mechanics, since it is not possible to define simultaneously and accurately the variables belonging to certain canonical pairs such as energy and time, momentum and position. A practical example of this limitation is the finite width, in terms of quantum energy or spectral frequency, of a spectrum line, arising from an excited state of finite life. This limit ($\sim 10^{-7}$ in many cases) is reached in the determination of a wavelength as a standard of length, but there is no reason why the length should not be defined to greater accuracy by the mean wavelength emitted from a great many atoms. In fact, the present experimental limit lies in finding a light source in which the systematic disturbing effect of fields external to the atoms is sufficiently reproducible.

Many measurements in practice are affected by fluctuations coarser than the quantum limit, and important examples are thermal fluctuations and the effect of the discrete (particulate) nature of matter and electricity.

Every degree of freedom determining the configuration of a system is associated with some energy, which appears as fluctuations of the appropriate coordinate. In principle, the statistics of these fluctuations must be calculated from quantum theory, but in many cases of practical importance, the average kinetic energy associated with a particular coordinate has the equipartition value $\frac{1}{2}kT$, and an approximate value of the uncertainty in amplitude can be calculated from this result. In practice, thermal fluctuations are most often encountered in electrical circuits because electronic devices have a high sensitivity extending over a wide frequency band. The fluctuation noise can now be put in the form of an alternating voltage in series with each resistor, distributed over all frequencies in accordance with the formula $E^2 d\nu = 4kTRd\nu$ (non-quantum approximation for frequencies such that $h\nu \ll kT$). Here E^2 denotes the distribution function for the squared voltage at frequency ν , T is the temperature and R the resistance. The effect on any circuit can be calculated from this result. It is apparent that the total noise voltage increases with the frequency bandwidth to which the system responds, so that it can be reduced in a system which has a long time of response, either because it responds only to low frequencies or because its response is narrowly selective. The phase-sensitive rectifier technique discussed above is a method of securing such a narrow response. Rapidly transient phenomena are, of course, excluded by any such system. It can be shown that if we can average our observations for a time τ , the relative accuracy varies as τ^{-1} . Thermal noise can be reduced in appropriate circumstances by reducing the temperature of the resistances involved, but the noise voltage distribution is then appreciably modified by quantum effects and does not disappear at absolute zero. It will be noted that thermal noise is explicitly connected with the resistive elements of a circuit. It can be reduced by using systems of amplification (parametric amplification) which do not in principle require resistive elements or by reducing the dissipative (resistive) element of a system by a properly phased feedback device.

In addition to the thermal noise, fluctuations can arise statistically from the passage of discrete electrons in a circuit, the full calculated value being observed in a saturated diode or in a photocell. The fluctuations are suppressed by the conduction mechanism of a metallic conductor or in part by the space charge of a triode or pentode, though in the pentode there is a new source of noise in the distribution of electrons between anode and screen. Noise of this kind (shot noise) is a major limitation in practical high-frequency measurements, but it is to some extent amenable to reduction by circuit techniques. The corresponding effect in a photocell is due to the finite rate of reception of quanta (N per second) and the corresponding rate of emission of electrons ($N\epsilon$ where ϵ is the cathode quantum efficiency), and it is inescapable. Furthermore, it is overlaid by the fluctuations of the "dark current" which may flow in the absence of incident quanta. This

dark current can be reduced by technical measures, and this must be done if the accuracy of the measurement of weak light is to be maximized. A further type of fluctuation, additional to thermal noise and shot noise, occurs in semiconductors and in the emission from most technical cathodes. It is not yet completely explained in theoretical terms; the squared distribution function is not uniform and increases rapidly at low frequencies; when this "excess noise" is present, it is desirable to "put the signal where the noise is not," i.e., to reduce the response of the measuring system at low frequencies and to make the signal appear at higher frequencies if the signal frequency can be chosen arbitrarily as in the case of a "chopper" detector or amplifier.

H. J. J. BRADDICK

Cross-references: ASTROMETRY; COSMIC RAYS; ELECTRICAL MEASUREMENTS; MAGNETOMETRY; NOISE; ACOUSTICAL; NUCLEAR INSTRUMENTS; OPTICAL INSTRUMENTS; PHOTOGRAPHY; PHOTOMETRY; THEORETICAL PHYSICS.

MECHANICS

The beginnings of the science of mechanics goes back to the Greeks inasmuch as they had derived some fundamental ideas on levers, vibrations of strings and on hydrostatics. However, not until the sixteenth century did mechanics make a real beginning whose progress has continued to the present day. Two great names are associated with the mechanics of the sixteenth and seventeenth centuries, namely Galileo Galilei in Italy and Sir Isaac Newton in England. Galileo (1564-1642) was essentially the one who insisted on the importance of observations of nature, of combining experiments and observations with careful thought, as opposed to the Aristotelians who at that time insisted that all knowledge came by thought alone. Many of the experiments that Galileo performed and the conclusions drawn from them are given in a book entitled "Two New Sciences," which, though written in a dialogue form, is still worthy of being read.

Following Galileo came another "natural philosopher," Sir Isaac Newton (1642-1727), who was one of the greatest of all time. It is interesting that the word "scientist" did not come into use until the middle of the nineteenth century, though the term "natural philosopher" is still used in some of the older universities. Newton's contributions in mathematics and physics were unique, for he invented the calculus and discovered the law of gravitation and the laws of motion. It is somewhat difficult to appreciate all that Newton did for the world. More than anyone else he ushered in the modern scientific era. The heavens were a region with its own particular laws, a region where speculation had relatively free reign. Newton changed all this. He applied to the planets the laws of motion, which were applicable to bodies on the earth, and he found he could account for the motion of the planets with remarkable accuracy.

So strong was the faith in the Newtonian principles that when the planet Uranus appeared to be deviating from its predicted orbit the assumption was made that there must be another planet as yet unknown. The position of this unknown planet was calculated and thus Neptune was found in 1846. Such was the success of the Newtonian conception of the universe.

Newton left the world a mechanical universe. Forces produced changes in motion between the sun and the planets just as they did between objects on the earth. It was the Newtonian idea that forces produced changes in motion in contradiction to the then current Aristotelean idea that force was required to keep motion going. Newton tried to be most careful in denying any "cause" in these laws of motion. He was trying to give a description of motion in terms of the laws. In his famous book "The Mathematical Principles of Natural Philosophy," usually referred to as "The Principia," Newton presents the scope of his work in mechanics, and what a grand scope it is. He not only discusses the motions of objects on the earth but the motion of the planets as well; the precession of the equinoxes, the tides and other problems in fluid motion.

The scientific caution which Newton showed in the Principia is illustrated by the following quotation: "For I here design only to give mathematical notion of those forces without considering their physical causes and seats." In another context he says: "You sometimes speak of GRAVITY as essential and inherent to matter. Pray do not ascribe that notion to me; for the cause of gravity is what I do not pretend to know."

Though there were no new physical principles, associated with mechanics, developed during the eighteenth and nineteenth centuries, there was much mathematical development. The Lagrange equations of motion are an elegant and different expression of the Newtonian equations of motion. Where Newton was concerned with forces, Lagrange was concerned with energy. In his "Mécanique Analytique" (1778), Lagrange showed how a mechanical problem could be solved on the basis of pure calculation without appealing to physical or geometrical considerations provided the kinetic and potential energies of the system were given in the appropriate form. Lagrange says in the preface "The reader will find no figures in the work. The methods which I set forth do not require either constructions or geometrical or mechanical reasonings; but only algebraic operations, subject to a regular and uniform rule of procedure." Considerable contributions to theoretical mechanics were also made by Sir William Rowan Hamilton (1805-1865) and C. G. J. Jacobi (1804-1851) together with a number of others: Gauss, Poisson, Poincaré and Mach.

Towards the middle of the last century, the principle of conservation of energy was firmly established on experimental grounds. This has been of enormous importance, for as different forms of energy have been recognized they have been seen to fit into a much wider perspective.

In the present century there has been the development of the theory of RELATIVITY by A. Einstein (1879-1955), important in the region of velocities close to the velocity of light, and of the QUANTUM MECHANICS applicable to the phenomena on the very small scale found in atoms.

It is obvious that motion is movement of some object relative to some other object. Newton recognized this in the Principia, and on the considerations of rotational motion, he found it necessary to postulate absolute space and absolute time. For motion at constant speed in a straight line it readily follows from Newton's laws that it is impossible to perform any physical experiment which can detect this motion. However, when the motion is accelerated, forces are required and the presence of the acceleration is readily determined. If a horizontal circular platform is turning with constant angular speed about a vertical axis through its center, then an observer on the moving platform experiences a centrifugal acceleration (see ROTATION—CIRCULAR MOTION). Also if he moves along a radius he experiences a sidewise acceleration called the CORIOLIS ACCELERATION. These accelerations are easily accounted for in terms of a rotating reference system. The earth is such a system and the centrifugal acceleration is shown in the systematic variation of the acceleration of gravity with latitude. For recognizable aspects of the Coriolis acceleration on the earth one has to examine large scale motions. The fact that the winds do not go directly from high pressure regions to low pressure ones in cyclones and anticyclones is readily explained by the Coriolis acceleration and force.

A major revision of thought came in mechanics in the present century when Einstein applied the theory of RELATIVITY to reference systems moving with constant velocity relative to one another. TIME and distance were not the same in the different systems though the speed of light was assumed to be constant in all the systems. This brought changes in mechanics which become appreciable when the speeds approach that of light, about 186,000 miles/sec. It is only in the realm of atomic and nuclear physics that speeds approaching this are encountered. In the macroscopic world of automobiles and airplanes the Newtonian mechanics is still applicable with considerable accuracy.

Mechanics has influenced many of the theories in other branches. It was a natural tendency to account for phenomena in mechanical terms whenever possible. Forces were experienced in electrical phenomena, and it was from the forces that entities such as electric charges were postulated. Vibratory motion, such as that undergone by an object hanging on a vertical spring, is used in explaining the scattering of light and x-rays.

There are two very important CONSERVATION LAWS in mechanics, namely conservation of momentum and conservation of energy. The expressions for these conservation laws, but not the principle of the laws, changes when the speeds of the objects approaches that of light. Then the

Einstein relativistic expressions must be used. For angular motion there is a corresponding law of conservation of angular momentum. Then there is the law of conservation of energy, a law which, when it includes mass m , as a form of energy E , by the relationship, $E = mc^2$, brings together two former conservation laws, namely the laws of conservation of mass and energy. The letter c indicates the velocity of light.

Thus we see that mechanics has a great theoretical structure and is also of great importance in the everyday world of machines and bridges. It has influenced all the sciences and today is as important as it ever was.

R. J. STEPHENSON

References

- Elementary: Stephenson, R. J., "Mechanics and Properties of Matter," New York, John Wiley & Sons, Inc., 1960.
Intermediate: Fowles, G. R., "Analytical Mechanics," New York, Holt, Rinehart and Winston, 1962.
Advanced: Goldstein, H., "Classical Mechanics," Reading, Mass. Addison-Wesley Co., 1950.

Cross-references: DYNAMICS, STATICS.

MEDICAL PHYSICS

That physics has an important place in medicine can scarcely be denied. A physician's first move in examining a patient is to measure his temperature, count his pulse, listen to his heart sounds and take his blood pressure. Only much later does the physician get around to chemical and laboratory tests. Yet every hospital of any stature has a laboratory or a department of clinical chemistry. Laboratories of clinical physics are virtually nonexistent. While physics plays a large role in medical diagnosis and treatment, physicists have largely neglected the field.

Some of the earliest applications of the principles of physics to problems in medicine were in the fields of optics and sound. An early contributor was H. L. F. von Helmholtz, a physician as well as a physicist. His work in physiological optics and that on the sensations of tone are considered classics. Even earlier, J. L. M. Poiseuille, a French physician and physicist, seeking a better understanding of the flow of blood, studied the flow of water in rigid tubes. His work not only contributed to physiology but also established an important relation in the physics of viscous fluids.

With the intensive development of the sciences of physics and medicine in the latter part of the nineteenth century, the two drew further apart. This period also saw rapid development in the science of physiology which is concerned not only with body chemistry but also with physical processes in the body. Clinical physiology abounds with such concepts as the pressure-velocity relationships in the flow of blood, the mechanics of the cardiac cycle, the work of breathing, gas exchange in the lungs, voltage gradient in cellular membranes, and cable properties of nerves, to

name but a few. These concepts have, of necessity, been worked out by scientists with training and experience in the basic biological and clinical procedures. Physicists have been inactive in the field and have made very little contribution to its development. But there is a growing awareness among physiologists of the importance of physical principles and the need for precise statement of physical law. An example of this conviction is the 18th edition of Howell's Textbook of Physiology which carries the title, "Medical Physiology and Biophysics."¹

A phenomenon of the mid-twentieth century has been the development of interdisciplinary fields of science. BIOPHYSICS combines the most fundamental of the biological and physical sciences. It has had an extremely rapid growth, with something like 30 to 40 university Departments of Biophysics in America alone. Its emphasis has been on the application of physical principles to all aspects of biology—cellular, botanical, zoological as well as clinical.

An even more recent phenomena has been the development of biomedical engineering. Its basis has been the application of the tremendous developments in electronics to medical measurements and instrumentation. In fact, the field is frequently referred to as biomedical electronics. Such recent developments as vector electrocardiography, implantable pacemakers for the heart, and intensive-care physiological monitors and recorders are examples of the impact of electronics on medicine. While these fields border on medical physics, none are concerned primarily with the application of physical principles to clinical problems. Yet they compete so effectively with medical physics that it is difficult to delineate the boundaries of the latter.

The discovery of x-RAYS by Roentgen in 1895 had an immediate impact upon medicine. Within a few months, the new rays were used both diagnostically and therapeutically. Indirectly, their application set the stage for the development of medical physics. Therapeutic application of x-rays raised questions concerning their quality and quantity—both of which are important in accurate dosimetry. Evaluation of early successes and failures indicated the importance of the proper distribution of dose between neoplasm and normal tissue. The physician turned to the physicist for assistance. The late Otto Glasser was one of the early radiological physicists; he and Fricke in 1924 constructed an air wall ionization chamber for the measurement of radiation dose.² Their construction eliminated some of the nonlinear effects due to quality, i.e., photon energy distribution, in the evaluation of biological response. Other early workers in America were Edith Quimby³ and the late G. Failla. In England, L. H. Gray and W. V. Mayneord were active. In 1936, Gray proposed the Bragg-Gray formula for determining the absolute amount of energy delivered to a medium from ionization measurements.⁴ The work of Fricke, Glasser, and Failla along with that of L. S. Taylor⁵ and others contributed to the establishment in 1928 by the

Second International Congress of Radiology of the roentgen as a unit of radiation dose based on the amount of ionization generated in a standard volume of air. The use of higher energies and ionizing radiations other than x-rays led during the 1950's to the abandonment of the roentgen as a unit of absorbed dose. Dissatisfaction with the roentgen was also due to a growing realization that biological response was more nearly related to the energy absorbed in a medium. The Bragg-Gray formula permitted the calculation of absorbed dose in a medium, and the work of J. S. Laughlin⁶ established the dosimetry of high-energy radiations in energy units by calorimetric methods. The International Commission on Radiological Units in 1956 adopted the *rad* as the unit of absorbed dose defining it as being equal to an absorption of 100 ergs per gram of material. The roentgen was retained as a unit of exposure dose for x-rays below 3 MeV, i.e., a measure of the intensity of a beam of x-rays to which a material might be exposed.

While radiological physics is clearly a part of the broader discipline of medical physics, it included in the early days practically all that was organized of the later subject. In 1943, Otto Glasser was persuaded to edit an encyclopedia on medical physics.⁷ It has since gone through three editions. It treats the physical aspects of the principal, medically oriented subjects including anatomy, dermatology, hematology, neurology, orthopedics, pathology, radiology and surgery, to name but a few, and provides a good survey of the scope of the field.

Until recently there has been no organization of workers in the field of medical physics. Radiological physicists were associate members of the North American Radiological Society, naturally dominated by radiologists. First in Britain, (the Hospital Physicists' Association) and later in America, specialty groups have been organized. The American Association of Physicists in Medicine brings together those physicists working in hospitals and medical schools and interested in an understanding of the physical side of medical problems. The membership has been largely drawn from those working in the area of radiological physics, but a growing interest in the broader area pertaining to all of medicine can be discerned. The organization cooperates with its British counterpart in the publication of a journal—*Physics in Medicine and Biology*. A further indication of the developing awareness of this field is the organization of the First International Conference on Medical Physics which was held at Harrogate, England in September 1965.

One last word about a related field: Radiation protection was in the early days a part of radiological physics. In America, L. S. Taylor was active for many years at the Bureau of Standards in setting up guidelines for protection from radiation. During World War II, the Manhattan Project required large numbers of workers in the field of protection, and the term **HEALTH PHYSICS** was introduced. Since the war, the field has grown with the growth of the area of atomic energy. The

Health Physics Society is a large and growing group with many local chapters and an international organization. The field seems, though, to be becoming more closely aligned with the area of public health than with clinical medicine.

That physics will play an increasingly important role in medicine cannot be challenged. But whether physicists will truly create a discipline of medical physics or whether that role will be played by physiologists, biophysicists, biomedical engineers or others remains to be seen.

LESTER S. SKAGGS

References

1. Ruch, T. C., and Fulton, J. F., Eds., "Medical Physiology and Biophysics," 18th edition, Philadelphia, W. B. Saunders, 1960, xxii, 1232pp.
2. Fricke, H., and Glasser, O., "Standardization of the Roentgen Ray Dose by Means of the Small Ionization Chambers," *Am.J. Roentgenol.* **13**, 462 (1925).
3. Glasser, O., Quimby, E. H., Taylor, L. S., and Weatherwax, J. L., "Physical Foundations of Radiology," Third edition, New York, Paul B. Hoeber, Inc., 1961, xi, 503pp.
4. Gray, L. H., "An Ionization Method for the Absolute Measurement of X-Ray Energy," *Proc. Roy. Soc., London Ser. A*, **156**, 578 (1936).
5. Taylor, L. S., and Singer, G., "An Improved Form of Standard Ionization Chamber," *J. Res., Natl. Bur. Std.* **5**, 507 (1930).
6. Genna, S., and Laughlin, J. S., "Absolute Calibration of Cobalt-60 Gamma Ray Beam," *Radiology* **65**, 394 (1955).
7. Glasser, O., Ed., "Medical Physics," Chicago, Year book Publishers, 1960, 3 vols., cxxii, 3725pp.

Cross-references. BIOPHYSICS, HEALTH PHYSICS, RADIOACTIVITY, X-RAYS.

MESON. See ELEMENTARY PARTICLES.

METALLURGY

The metallurgical industry is one of the oldest of the arts, but one of the youngest of the subjects to be investigated systematically and considered analytically in the tradition of the pure sciences. It is only in comparatively recent times that any fundamental work has been carried out on metals and alloys, but there are now well-established and rapidly growing branches of science which are related to the metallurgical industry.

Process Metallurgy. Process metallurgy, or the science of extracting metals from their ores, is broadly divided into two groups.

Ferrous. This branch is concerned with the production of *iron* (normally from iron ore, with coke and limestone in a blast furnace) and its subsequent refining to *steel*, by oxidizing the impurities either in an "open hearth" furnace by means of an appropriate slag on the surface or in a "converter" by blowing air through the molten iron. The most striking recent development in this field has been the increasing use of pure

gaseous oxygen in steelmaking, with a resultant improvement in efficiency, rate of production, and quality of product.

Non-ferrous. This is concerned with the production of the remaining metals. Those manufactured in greatest quantity include aluminium, copper, nickel, zinc, magnesium and tin, with titanium being an important newcomer in view of its low density, high melting point (1670 °C) and resistance to corrosion. The precious metals, and the "refractory metals" of very high melting point (e.g., tungsten and molybdenum) are other important families.

Shaping of Metals. This may be carried out in three main ways:

Casting. Although most metals are initially cast into *ingots*, which may be subsequently forged to the desired shape, many alloys are designed to be cast into their final shape by pouring the molten alloy into an appropriate mold. These may be sand molds if only a small number of objects are required, and very massive castings (e.g., over 100 tons) may also be produced in this way. A permanent mold, or die casting, is employed if large numbers of the object are required (particularly in alloys of low melting point, such as zinc-based alloys) and high dimensional accuracy can be achieved by these means.

Forging. This entails shaping of the metal by rolling, pressing, hammering, etc., and may be carried out at high temperatures, when the metal is soft (hot working), or at lower temperatures (cold working), where the deformation leads to progressive hardening of the metal (work hardening). In contrast with castings, forgings usually exhibit differing physical and mechanical properties in different directions, due to the directional nature of the shaping operation.

Much modern research in physical metallurgy is concerned with investigating the plastic flow and work hardening behavior of metals and alloys. Metal crystals yield plastically at stresses several orders of magnitude lower than the theoretical value for the deformation of perfect crystals: this discrepancy is accounted for by the presence of imperfections known as "dislocation lines" within the crystals. Plastic flow takes place in metal crystals by "slip" or "glide" in definite crystallographic directions on certain crystal planes, due to the movement of dislocation lines under the applied stress. Dislocations multiply and entangle as deformation proceeds, thus making further flow increasingly difficult (work hardening)—the density of dislocations rising from about $10^7/\text{cm}^2$ in soft (annealed) metal to about $10^{12}/\text{cm}^2$ in work-hardened material. Such crystal defects can be studied by x-ray diffraction and by modern electron micrographic means.

Powder Metallurgy. This is a method of shaping by pressing finely powdered metal into an appropriately shaped die. The "green compact" thereby produced is of low strength and is subsequently heated in an inert atmosphere ("sintered"); the pressing and sintering may be repeated until strong, dense products are obtained. The technology was first developed for metals which were

of too high a melting point for conventional casting and forging methods, and tungsten lamp filaments were first produced by this means. Other refractory metals and hard metal-cutting alloys may be thus shaped, and some magnetic and other special alloys are prepared in this way by suitable blending of powders, which avoids any contamination that may be associated with the melting process. The pressing and sintering conditions may be arranged to leave some residual porosity in the structure of, for example, bronze bearing alloys. The pores are filled with oil, thus producing the so-called oil-less bearings which can operate without further lubrication.

Joining. Rivetting, soldering and brazing (in which metal components are joined by means of a layer of a low-melting alloy), and welding are the important methods of joining metals. *Weldability* is often the critical factor in the selection of an alloy for a given purpose, since the metallurgical changes produced by localized heating are often associated with the development of deleterious properties at, or adjacent to, the weld.

Alloy Constitution. Phase equilibria in alloy systems are represented on *phase diagrams*, which are experimentally established by thermal analysis, microscopical, and x-ray diffraction methods. Phase diagrams are invaluable in the interpretation of the structures of alloys observed under the microscope.

The microstructure of an alloy (and hence its properties) will be determined not only by its composition, but also by its thermal and mechanical history. Of particular importance is the metallurgical control of the mechanical properties of an alloy by *heat-treatment*, which affects the distribution of the phases present. Hardness, for example, will depend upon the state of deformation (i.e., the density of dislocations) and upon the composition of the alloy. Pure metal crystals can be hardened by other atoms in solid solution (solute hardening) as well as by finely dispersed particles of a hard second phase (precipitation, or dispersion hardening) which are effective in impeding the motion of dislocations when the crystal is stressed. The relationship between the microstructure and properties of metals and alloys is thus of fundamental importance and is a field of intense scientific activity.

Although many common alloys were not developed scientifically, there is a considerable theory of alloy developing, springing from empirical rules and principles (notably those due to W. Hume-Rothery) which have generalized the facts and enabled predictions to be made. The early theories of the metallic state, due to Drude and Lorentz, and later to Sommerfeld were developed and discussed by N. F. Mott and H. Jones in their book "The Theory of the Properties of Metals and Alloys." Although during the last decade the early electron theory of copper alloys has gradually been destroyed by recent theoretical and experimental work, a great increase in our knowledge of transition metals and their alloys has taken place, and some signs of general principles have begun to appear, although there is yet

little theoretical knowledge enabling one to calculate properties or structures of alloys from fundamental principles.

The Effect of Environment Upon the Behavior of Metals.

Low Temperature. Some metals and alloys exhibit a spectacular change in mechanical behavior with decrease in temperature. Many metals of body-centered cubic crystal symmetry (e.g., iron and mild steel) which are tough and ductile at ordinary temperature become completely brittle at subzero temperatures, the actual transition temperature depending upon the metallurgical condition of the alloy, the state of stress, and rate of deformation. Some metals of hexagonal symmetry (e.g., zinc) exhibit this effect, but metals of face-centered cubic symmetry (e.g., copper) remain ductile to the lowest temperatures. This transition in behavior is clearly of critical importance in the selection of materials for low-temperature application.

High Temperature. Apart from problems of oxidation (discussed below), metals tend to deform under constant stress at elevated temperatures (the deformation is known as "creep"), and creep-resistant alloys are designed to provide strength at high temperatures. These are essentially alloys in a state of high thermodynamic stability, usually containing finely dispersed particles of a hard second phase which impede the movement of dislocations.

Fatigue. Metals break under oscillating stresses whose maximum value is smaller than that required to cause rupture in a static test, although many ferrous alloys show a "fatigue limit," or stress below which such fracture never occurs, however great the number of cycles of application. The phenomenon is associated with the nucleation of submicroscopic surface cracks in the fatigued component early in its life, which initially grow very slowly. Eventually a crack grows until the effective cross section of the piece is reduced to such a value that the applied stress cannot be supported, and rapid failure occurs.

Oxidation and Corrosion. Apart from the noble metals, when metals are heated in air, they owe their oxidation resistance to the presence of impervious oxide films of their surface, and those which develop porous oxides (e.g., the refractory metals tungsten and molybdenum) oxidize very rapidly at high temperatures. Oxidation resistant alloys are designed to maintain a protective film under these conditions.

Corrosion occurs under conditions of high humidity or immersion in aqueous media. The phenomenon can be interpreted electrochemically—local anodes form at the region of metal dissolution, and local cathodes form where the electrons are discharged. "Galvanic corrosion" is encountered where dissimilar metals are in electrical contact under these conditions. Of particular importance is the *conjoint* action of stress and corrosion, where "stress corrosion" or (under fluctuating stresses) "corrosion fatigue" cracking may be encountered, in situations where no failure would occur under the action of the stress

or the corrosive environment applied separately. Electrochemical principles are applied in the protection against corrosion.

JOHN W. MARTIN

References

Metallurgical Data

- "Metals Handbook," American Society for Metals, Cleveland, Ohio.
- Smithells, C. J., "Metals Reference Book," London, Thornton Butterworth, Ltd., 1962.
- Hansen, M., "Constitution of Binary Alloys," New York, McGraw-Hill Book Co., 1958.

General Reading

- Street, A., and Alexander, W. O., "Metals in the Service of Man," London, Pelican.
- Guy, A. G., "Elements of Physical Metallurgy," Reading, Mass., Addison Wesley, 1959.
- Hume-Rothery, W., "Electrons, Atoms, Metals and Alloys," New York, Dover, 1963.
- Evans, U. R., "An Introduction to Metallic Corrosion," London, Edward Arnold, 1963.

Cross-references: CREEP AND FATIGUE, CRYSTALS AND CRYSTALLOGRAPHY, ELECTROCHEMISTRY, SOLID-STATE PHYSICS

METEOROLOGY

The theme of meteorology over the past two decades has been expansion; expansion in all its aspects; expansion at accelerating pace. Meteorology has moved into an era of quantization and specialization. The reasons for this growth are not hard to find. The world is making ever greater use of its atmosphere and must know more about its nature. At the same time, the modern electronic tools for probing the atmosphere in depth have become available. As a result there are more data, more interest, more research, and more services than ever before.

Traditionally meteorological data were gathered by surface observers. Thirty years ago regular observations of the upper atmosphere began on a regular basis by means of balloon soundings using radio telemetry to transmit the data to the ground station. This proved to be the beginning of a veritable explosion in the sampling of the atmosphere. Instrumentation and balloon quality have steadily improved, but the network stations are now approaching the altitude limit of 32 km. Beyond this level, the rocket has replaced the balloon as the vehicle for sampling to about 60 km. Density, temperature and wind are being measured by a small but growing network. Beyond 60 km, rocket data are limited by lack of instruments capable of measuring under these conditions. This will soon be overcome.

At a still higher level, the satellite operates as an observing platform. Sensors on this platform measure various types of radiation. From radiations in the visible bands come the familiar and impressive pictures of clouds and terrain. High-resolution infrared radiation is providing pictures

on the dark side of the world as well as measurements of the temperature on the earth, clouds, and atmosphere. Other bands tell how much radiation is being removed by water vapor and other absorbers and, hence, provide a measure of the quantities and distributions of these constituents. The satellite samples equally well over difficult terrain and populated areas, and it promises to provide answers to one of the scientists' great problems—that of obtaining adequate data on a global scale. It also provides a direct method of obtaining the radiational balance of the earth and atmosphere, a factor of fundamental importance in the energy budget of the earth-atmosphere system and a vital parameter in the general circulation and in long-range forecasting.

Meteorology requires data over wide areas as well as to great heights and here again progress has been great. Observational instructions have been standardized on a global basis and data are freely exchanged. Advances in speed of transmission have increased the volume of data exchanged, and the evidence is that the speed will increase further by many orders of magnitude. Processed data are being exchanged also in numerical, pictorial, and digital forms. In addition to the original radio and teletype, there is the facsimile for visual presentations and the communications satellite which will be important in future global networks.

The tools of measurement have undergone equally impressive development. The radiosonde is no longer restricted to measurement of the traditional pressure, temperature, humidity, and wind. It may also measure radiational flux, ozone distribution, and other factors in the upper air. Radar has come of age as a tool of great power in probing the secrets of cloud and precipitation physics. Micrometeorological systems consisting of thermometers and anemometers measure the parameters important for analysis of turbulent flow and diffusion. They also exemplify the growing tendency to create systems which feed out processed rather than raw data, in this case in the form of statistical data in digitized form.

With the vastly increased capacity for obtaining data has come an increased capacity for digesting it. Data which may have future value are stored, often in the form of punched cards. However with increasing masses of data even the storage of punched cards is a problem, and even more concise forms are being developed. Data, particularly analyzed data, are being stored in digitized form on magnetic tapes. This makes them readily available for analysis on an electronic computer. With this sort of capability, the whole approach to research is undergoing change. In many areas, the analysis by electronic computer has become the normal approach. Synoptic analysis and prediction is rapidly moving into that class.

As one might expect with the avalanche of new data, research into the structure and behavior of the atmosphere has been experiencing a corresponding development. This is exemplified

by the growth of institutions for research and teaching in meteorology. Thirty years ago, there were in the world only a few universities active in meteorology. Today there are 20 to 30 in the United States alone, and there are many more in certain specialized areas of the science. This period has seen the manyfold increase in direct participation by the governmental agencies in meteorological research. Journals to publish the resulting research have matured and multiplied. The private meteorological consultant has come into his own. Many have built medium-sized firms providing a large amount of specialized service including research under contract to private companies as well as to the government.

Thirty years ago, meteorology largely consisted of a semiquantitative science concerned with the synoptic scale motions of the atmosphere, those of importance to forecasting for the next 48 hours, and certain basic physical principles. This picture has changed completely in the intervening years. Gone is the simple flow concept of the atmosphere, and in its place is an atmosphere of almost unbelievable complexity. In place of the synoptic scale pattern of wave motion in the atmosphere, we have a hierarchy of scales. These include the hemispheric long waves, or Rossby waves, in the westerly flow having wavelengths of the order of 5000 km, and the synoptic scale of wavelength of the order of 1500 km. Then follows the mesoscale, phenomena of the order of 10 to 50 km, and the microscale which involves motions of the order of a few meters. According to Kolmogoroff, the driving energy of the atmosphere is transferred from one scale of motion to another of lower scale. Large-scale motions break down into smaller motions which themselves break down into still smaller motions. Thus energy cascades down the scale, the limit being reached when energy is absorbed into motions of the molecules showing up as heat.

The middle-latitude westerly stream of air, formerly thought to be a broad steady flow, is organized into two or more jet streams; narrow streams of air concentrated at tropopause level and extending around the world with speeds up to 100 m/sec. The jet streams generally meander northward and southward outlining a more or less regular wave pattern of the order of 90° longitude wavelength. These Rossby waves move only slowly. Ridges in the system correspond to warmer than average air and troughs to colder. Thus these systems are the key to extended range forecasting, forecasting of the order of 5 days in advance.

The synoptic scale motions were, of course, much studied in earlier years. Nevertheless, with better upper air data and knowledge of motions in other scales, advances are still being made. Perhaps the most significant advance in both the hemispheric scale and the synoptic scale has been made possible by the electronic computer. The equations of motion of the atmosphere have been known since the early days of the science. However, these equations are nonlinear. A functional solution is not possible. A numerical solution is

possible under simplifying assumptions but only if facilities are available for extensive calculation. The electronic computer has made this possible. Research over the past 15 years has shown which assumptions are necessary to produce representative mathematical models of the atmosphere amenable to solution. As it turns out, quite useful results can be obtained by a very rigid restriction, the assumption of a barotropic atmosphere. In such an atmosphere, all levels behave in exactly the same manner so that this model can only approximate the real atmosphere at one level. This level is approximately 4.5 km altitude. It was many years before baroclinic models, models without this restriction, produced as good results. This point has now been reached, and present models are able to predict the effect of various influences on the development of cyclones and Rossby waves.

Insofar as these scales of motion are concerned, the computer has become the laboratory of the meteorologist in the truest sense. By altering the parameters in his models, the meteorologist can test the effects of various influences on the atmosphere. The newer models incorporate the effects of solar heating, radiational cooling, formation of clouds and precipitation, as well as the peculiarities of the earth's surface. Much work has been done, but this phase of the science is poised for great advance.

Spectacular changes, based on this work, are taking place in weather forecasting. Large-scale motions are being predicted through the use of electronic computers. Data from the teletype circuits are fed directly to computers which check them for consistency, analyze them for values of the variables at fixed grid points, predict the flow patterns, and draw the synoptic weather map. At the moment, the most success by this method is found in the middle troposphere, but it is only a matter of time till all levels can be accurately forecast by these means.

The mesoscale studies deal with phenomena such as thunderstorms and other small scale circulation systems. This scale of phenomena is also related to very short-range forecasting which is important for certain activities. An example is the forecasting for landing of supersonic aircraft in critical conditions. Such aircraft cannot fly long at low altitudes, and it is essential that the pilot have an accurate forecast for the next twenty to thirty minutes before descending. In the forecasting business, specialists in mesoscale predictions are being developed to work as a team with the specialists in the hemispheric and synoptic scales. This permits a better all-round use of the data. At all levels, the use of the computer will increase as research points the way toward more quantitative methods.

Molecular diffusion is such a slow process that it is of only minor importance in the atmosphere, but diffusion by means of turbulent eddies, eddy diffusion, is a phenomenon of extreme importance to modern man. The statistical description of such diffusive processes has been developed to a high degree. The micro-

meteorologist has become a key figure in man's fight against pollution of the atmosphere and its harmful effects. Besides the obvious pollutants of smoke and noxious gases there is pollution due to the increased use of atomic energy as a source of power. In this sensitive area, an accurate knowledge of the diffusion characteristics of the neighboring region is essential.

One of the great new fields of meteorological research is that of CLOUD PHYSICS. The major tool is the radar, but data are also gathered by other means such as instrumental aircraft and conventional ground instrumentation forming a very fine network. Clouds form when moist air rises, expands and cools. While condensation nuclei are necessary before droplets will form, the atmosphere always has an ample supply of these so that clouds always form when saturation is reached. It has been found also that there is no fixed freezing point for water in the atmosphere and indeed condensation takes place in the liquid state at temperatures substantially below 0°C. The small droplets so formed have proved to be exceedingly stable, and one of the major problems of meteorology has been to learn more of the mechanisms which permit these droplets to grow into large ones. It has long been known that if some droplets freeze, there is the possibility of a distillation from the small droplets to the ice crystals causing the latter to grow. In middle latitudes, this is believed to be the major cause of growth of droplets to raindrop size. There are always a few nuclei in the atmosphere suitable for use as crystallization nuclei, but there are insufficient to cause a large percentage of the droplets to crystallize. Much research has been done to try to improve the richness of the atmosphere with respect to ice nuclei. Silver iodide has proved to be useful, because of its similar structure to that of the ice crystal, and is active at temperatures of about -5°C. Many experiments have been made using silver iodide as a seeding agent to supply additional nuclei to the atmosphere and to increase rainfall. Because of the natural variability of rainfall, it requires a very carefully designed experiment to permit meaningful statistical evaluation of such experiments. Even now it is uncertain that rainfall is actually increased significantly by such techniques except in regions where there is forced uplift of the air as in mountainous areas. Nevertheless, because of the great value that can be attached to even a modest increase in some instances, the commercial cloud-seeding operator has been able to find much business. The other major mechanism for growth of droplets is that of coalescence. This mechanism depends on the production of a few large drops, generally by giant condensation nuclei of hygroscopic salt, which then sweep out swaths of smaller droplets due to differential motion and momentum. This mechanism, important in the tropics, seems to be less amenable to adjustment by man.

Some of the more exciting areas of new knowledge lie in the higher atmosphere. The atmosphere seems to divide naturally into various

layers. The lowest of these is the troposphere where convection rules. The temperature drops off steadily to about 9 or 10 km. The top of this layer is known as the tropopause. The temperature in the lower stratosphere is approximately constant with height but soon begins to increase until it reaches a peak at about 50 km with temperatures comparable to those at the earth's surface. This level is known as the stratopause. The temperature then drops off with height through the mesosphere to about 90 km, a level known as the mesopause.

With increasing amounts of data, what originally seemed to be a rather tranquil picture of the upper atmosphere has undergone some change. We find that the arctic stratosphere generates a rather strong westerly jet stream at about 25 km altitude in wintertime. The winter arctic stratosphere is subject to more or less sharp warmings of substantial magnitude each year in the period between about January and March. This sudden warming marks the end of the stratospheric winter which then changes over to a summer regime. There is also a noticeable 26-month cycle in the tropical stratospheric wind, the direction changing from easterly to westerly about every 26 months. A similar tendency, but much more subdued, has been found at higher latitudes. There is evidence that this cycle affects the stratospheric warmings causing them to develop in Europe first one year and North America the next.

The whole field of services has grown as more knowledge is gained. It is possible here to give only some instances, but it should be realized that this development of widely diversified services based on meteorology is one of the great developments of the last 15 years.

D. P. MCINTYRE

Cross-references: CLOUD PHYSICS, COMPUTERS, HEAT, PLANETARY ATMOSPHERES, TELEMETRY, TEMPERATURE AND THERMOMETRY.

MICHELSON-MORLEY EXPERIMENT

Introduction. The revival and development of the wave theory of light at the beginning of the nineteenth century, principally through the contributions of Young and Fresnel, posed a problem which proved to be of major interest for physics throughout the entire century. The question concerned the nature of the medium in which light is propagated. This medium was called the "aether" and an enormous amount of experimental and theoretical work was expanded in efforts to determine its properties. On the experimental side, a long series of electrical and optical investigations were carried out attempting to measure the motion of the earth through the ether medium. For many years, the experimental precision permitted measurements only to the first power of the ratio of the speed of the earth in its orbit to the speed of light ($v/c \approx 10^{-8}$), and these

"first-order experiments" uniformly gave null results. It became the accepted view that the earth's motion through the ether could not be detected by laboratory experiments of this sensitivity. With the development of Maxwell's electromagnetic theory of light, and especially with its extensions by Lorentz in his electron theory, theoretical explanations for the null results obtained in the early ether drift experiments were provided. These results were in harmony with the Galilean-Newtonian principle of relativity in mechanics, which explains why the essential features of all uniform motions are independent of the frame of reference in which they are observed. In Maxwell's electromagnetic theory, however, the situation was altered when quantities of the second order in (v/c) were considered. According to the Maxwell theory, effects depending on $(v/c)^2$ should have been detectable in optical and electrical experiments. The presence of these "second-order effects" would indicate a preferred reference frame for the phenomena in which the ether would be at rest. At first, this feature of Maxwell's theory implying observable ether drift effects of the second order in (v/c) raised a purely hypothetical question, since the accuracy needed for such experiments was a part in a hundred million, and no experimental techniques then known could attain this sensitivity.

Michelson pondered this problem and it led him to invent the Michelson interferometer, which was capable of measurements of the required sensitivity, and to plan the ether drift experiment which he carried to completion in collaboration with Edward W. Morley at Cleveland in 1887. This famous optical interference experiment was devised to measure the motion of the earth through the ether medium by means of an extremely sensitive comparison of the velocity of light traveling in two mutually perpendicular directions. The experiment, when completed in 1887, gave a most convincing null result and proved to be the culmination of the long nineteenth century search for the ether. At that time, the definitive null result of the Michelson-Morley experiment was a most disconcerting finding for theoretical physics, and indeed for many years repetitions of this experiment and related ones were performed with the hope of finding positive experimental evidence for the earth's motion through the ether. These later experiments, however, have all been shown to be consistent with the original null result obtained by Michelson and Morley. In the years following 1887, their experiment led to extensive and revolutionary developments in theoretical physics, and proved to be a major incentive for the work of FitzGerald, Lorentz, Larmor, Poincaré, and others, leading finally in 1905 to the special theory of relativity of Albert Einstein.

The optical paths in the Michelson-Morley interferometer are shown in plan in Fig. 1. Light from a is divided into two coherent beams at the half-reflecting, half-transmitting rear surface of the optical flat b. These two beams travel at 90° to each other and are multiply reflected by

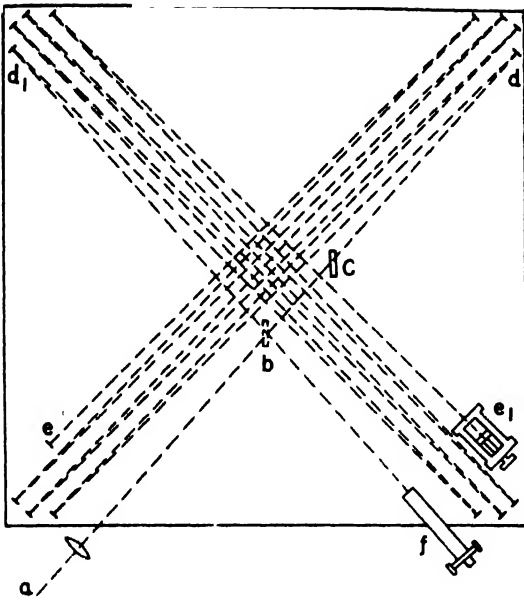


FIG. 1 Optical paths in the Michelson-Morley interferometer.

two systems of mirrors $d - e$ and $d_1 - e_1$. On returning to b part of the light from $e - d$ is reflected into the telescope at f , and light from $e_1 - d_1$ is also transmitted to f . These two coherent beams of light produce interference fringes. These are formed in white light only when the optical paths in the two arms are exactly equal, a condition produced by moving the mirror at e_1 by a micrometer. c is an optical compensating plate. The effective optical path length of each arm of the

apparatus was increased to 1100 cm by the repeated reflections from the mirror system.

Figure 2 is a perspective drawing of the Michelson-Morley interferometer showing the optical system mounted on a 5 foot square sandstone slab. The slab is supported on the annular wooden float, which in turn fitted into the annular cast-iron trough containing mercury which floated the apparatus. On the outside of this tank can be seen some of the numbers 1 to 16 used to locate the position of the interferometer in azimuth. The trough was mounted on a brick pier which in turn was supported by a special concrete base. The height of the apparatus was such that the telescope was at eye level to permit convenient observation of the fringes when the instrument was rotating in the mercury. While observations were being made, the optical parts were covered with a wooden box to reduce air currents and temperature fluctuations.

This arrangement permitted the interferometer to be continuously rotated in the horizontal plane so that observations of the interference fringes could be made at all azimuths with respect to the earth's orbital velocity through space. When set in motion, the interferometer would rotate slowly (about once in 6 minutes) for hours at a time. No starting and stopping was necessary, and the motion was so slow that accurate readings of fringe positions could easily be made while the apparatus rotated.

The experiment to observe "the relative motion of the earth and the luminiferous ether" for which this instrument was devised, was planned by Michelson and Morley as follows. When the interferometer is oriented as in Fig. 3 with the arm L_1 parallel to the direction of the earth's velocity v in space, the time required for light to

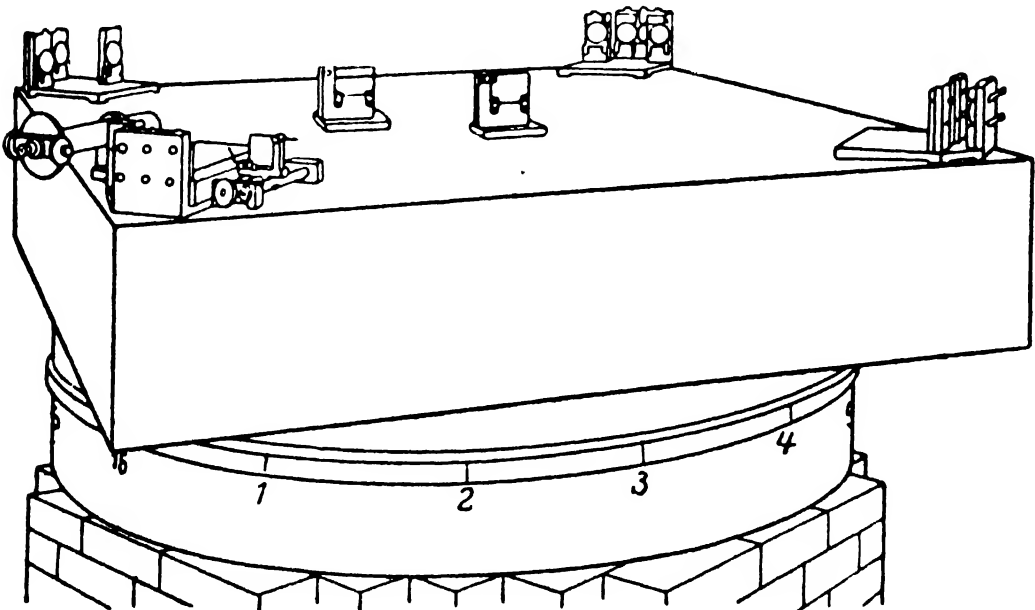


FIG. 2. Michelson-Morley interferometer used at Cleveland in 1887.

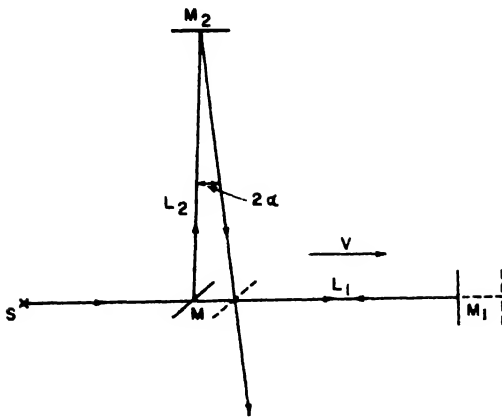


FIG. 3. The Michelson-Morley experiment.

travel from M to M₁ and return to M in its new position is,

$$t_1^{(1)} = \frac{L_1}{c-v} + \frac{L_1}{c+v} = \frac{2L_1}{c} \frac{1}{1-\beta^2} \quad \left(\beta = \frac{v}{c} \right)$$

The time for light to make the journey to and from the mirror M₂ in the other interferometer arm L₂ is,

$$t_2^{(1)} = [2L_2(1 + \tan^2 \alpha)^{1/2}] / c$$

and since $\tan^2 \alpha = v^2 / (c^2 - v^2)$

$$t_2^{(1)} = \frac{2L_2}{c} \frac{1}{(1-\beta^2)^{1/2}}$$

When the interferometer is rotated through 90° in the horizontal plane so that the arm L₂ is parallel to *v*, the corresponding times are,

$$t_2^{(2)} = \frac{2L_2}{c} \frac{1}{1-\beta^2}$$

$$t_1^{(2)} = \frac{2L_1}{c} \frac{1}{(1-\beta^2)^{1/2}}$$

Thus, the total phase shift (in time) between the two light beams expected on the ether theory for a rotation of the interferometer through 90° is,

$$\begin{aligned} \Delta t = & \frac{2L_1}{c} \left[\frac{1}{1-\beta^2} - \frac{1}{(1-\beta^2)^{1/2}} \right] \\ & + \frac{2L_2}{c} \left[\frac{1}{1-\beta^2} - \frac{1}{(1-\beta^2)^{1/2}} \right] \\ = & \frac{2(L_1 + L_2)}{c} \left[\frac{1}{1-\beta^2} - \frac{1}{(1-\beta^2)^{1/2}} \right] \end{aligned}$$

For equal interferometer arms, as used in this experiment,

$$L_1 = L_2 = L, \text{ and, since } \beta \ll 1,$$

$$\Delta t \approx \frac{2L}{c} \beta^2$$

The observations give the positions of the fringes, rather than times, so the quantity of

importance for the experiment is the change in optical path in the two arms of the interferometer.

$$\Delta = c \Delta t = 2L(v/c)^2$$

This is the quantity of second order in (*v*/*c*) referred to above.

With the Michelson-Morley interferometer, the magnitude of the expected shift of the white-light interference pattern was 0.4 of a fringe as the instrument was rotated through an angle of 90° in the horizontal plane. Michelson and Morley felt completely confident that fringe shifts of this order of magnitude could be determined with high precision.

In July of 1887, Michelson and Morley were able to make their definitive observations. The experiments which gave their final measurements were conducted at noon and during the evening of the days of July 8, 9, 11, 12 of 1887. Instead of the expected shift of 0.4 of a fringe they found "that if there is any displacement due to the relative motion of the earth and the luminiferous ether, this cannot be much greater than 0.01 of the distance between the fringes."

The result of the Michelson-Morley experiment has always been accepted as definitive and formed an essential base for the long train of theoretical developments that finally culminated in the special theory of relativity. The first important suggestion advanced to explain the null result of Michelson and Morley was G. F. Fitzgerald's hypothesis that the length of the interferometer is contracted in the direction of its motion through the ether by the exact amount necessary to compensate for the increased time needed by the light signal in its to-and-fro path. This contraction hypothesis was made quantitative by H. A. Lorentz in further development of his electron theory in which he introduced the formalism which has since been known as the "Lorentz transformation" for the analysis of relative motions.

H. Poincaré also contributed greatly to both the philosophical and mathematical developments of the theory. As early as 1899, he asserted that the result of Michelson and Morley should be generalized to a doctrine that absolute motion is in principle not detectable by laboratory experiments of any kind. Poincaré further elaborated his ideas in 1900 and in 1904 and gave to his generalization the name "the principle of relativity." He also completed the theory of Lorentz and it was he who named the essential transformation "the Lorentz transformation."

In 1905 Einstein published his famous paper on the "Electrodynamics of Moving Bodies" in which he developed the special theory of relativity from two postulates: (1) the principle of relativity was accepted as the impossibility of detecting uniform motion by laboratory experiments, and (2) the constancy of the speed of light was generalized to a postulate that light is always propagated in empty space with a velocity independent of the motion of the source. Both postulates have a close relationship to the Michelson-Morley experiment, which Einstein knew through

his study of the work of Lorentz. Einstein's paper is generally considered as the definitive exposition of the special relativity principle, and the climax of the century-long developments which had begun with Young and Fresnel to explain the electrical and optical properties of moving bodies. It has since become a major factor in the modern development of both classical and quantum physics.

R. S. SHANKLAND

Cross-references: INTERFERENCE AND INTERFEROMETRY; LIGHT; OPTICS, GEOMETRICAL; OPTICS, PHYSICAL; RELATIVITY.

MICROSCOPE

The word "microscope" comes from a Greek word that means "to view small." Before the microscope could come into existence, the LENS had to be developed. Early knowledge of lenses is very obscure, but we do know that the Chinese porcelain vases had figures on them wearing glasses. These vases date back to 1000 B.C. The Assyrian lens of 700 B.C. was found to be double convex. Sir Austen Henry Layard discovered a plano-convex lens of rock crystal at the Nineveh excavation. The Greeks and Romans wrote about the burning properties of glass lenses as well as the use of glass globules filled with water to aid seeing small objects. Roger Bacon wrote about the magnification power of lenses to aid elderly persons in reading. After his book "Opus Majus" was published in 1266 A.D., nearly two hundred years passed before anything was written discussing advancements in optics. In 1542, the first references about telescopes appeared, and these were written by Nicholas Copernicus of Poland, who discovered that the sun is the center of our universe and that other universes exist.

In Middleburg, Holland, about 1590, the story of the compound microscope begins. Zacharias Janssen developed, by accident, a microscope which was 1 inch in diameter and 6 feet long. This discovery started the parade of inventors and improvers of the microscope, too numerous to mention in their entirety in this article.

Galileo in 1610 introduced a microscope similar to Janssen's but with one major improvement. The Galileo microscope could be focused by means of screw threads on the body and mount. Fontana, Drebbel and Kepler also made microscopes in this early period.

The early microscope lenses had been formed on the end of glass rods by heating and pressing into a given shape. Campani was the first to design and grind lenses which had curves that could be reproduced. Anton van Leeuwenhoek studied all types of specimens under a simple microscope with ground lenses of his own design. He made in excess of 100 microscopes with single lenses, some of which were $\frac{1}{8}$ inch in diameter and had very strong curvature. With these instruments, he discovered bacteria and the existence of corpuscles in the blood.

Robert Hooke, in the last quarter of the 1600's, made a compound microscope that was easy to use. Hooke used a doublet eyepiece from Christian Huyghens to make up the telescope portion (eyepiece) of his compound microscope. The Royal Microscopical Society published Robert Hooke's "Micrographia" in which he describes tissues, blood vessels, textiles, papers, sugars and salt crystals. In about 1750, another Englishman, by the name of John Dollond, improved lenses by devising an achromatic system compounded from both hard and soft glass. In 1759, Dollond succeeded in making an achromatic lens. However, this type of lens was not used in a microscope until 1825. In 1854, the first steps toward standardization of the microscope took place with the "X" designation for magnification based on 250 mm. In 1873, Ernst Abbe from the University of Jena wrote "Theory of Microscope Image Formation" and "Sub-Stage Illumination Apparatus." If the theories discussed in Abbe's book had not been followed, today's microscopes would form very poor, inferior images.

Objective Lenses. The use of simple lenses with high curves for high magnification presents problems that make their use impractical. As the curve of the lens increases, the focal length becomes shorter and more difficult to use. An example of this is the Leeuwenhoek microscope in which one's eye had to be placed on top of the lens in order to see the magnified specimen. A microscope is nothing more than an objective lens magnifying a specimen, the image of which is being viewed by a simple telescope at some convenient distance from the objective lens (Fig. 1).

The aberration of lenses is always a problem to the designers of objectives. The two most common problems are color aberration (chromatic) and distortion of shape aberration (spheric). If a beam of white light passes from a medium of one density into a medium of another density such as air to glass, the beam is bent. As white light is made up of different colors, each wavelength refracts a different amount, and thus we have a spread of the colors called dispersion (Fig. 2). The higher the index of refraction the greater the dispersion. Figures 3 and 4 show the reasons that aberrations appear in lenses. A simple, and many times the most economical, practical way to correct spherical aberration is to place a diaphragm in front of the lens so as to cut off the thin edges which bend the light more (Fig. 4 and 5). To correct chromatic aberrations various types of glass must be selected for index of refraction and dispersion; they should then be combined to have as many colors as possible come to focus at the same point.

The objective is the most important single factor of the microscope, for through its power to resolve minute structure, we see small objects crisp and clear. Magnification, although it is of secondary importance to the resolving power of the objective, is absolutely essential. In order to have high resolution, a lens must be the best possible compromise on the correction of the

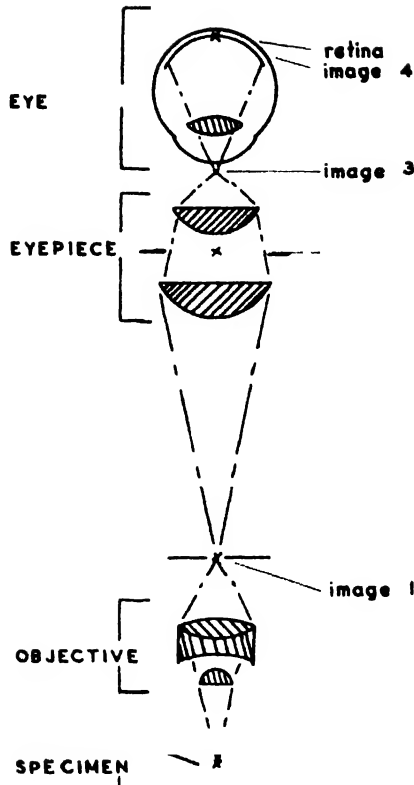


FIG. 1. Compound microscope showing optics and image position.

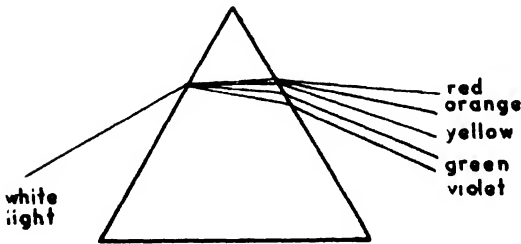


FIG. 2. Bending illustrates refraction; spreading illustrates dispersion.

various aberrations such as spherical, coma, and chromatic. The matter of resolution in actual practice is based not only on the correction of the lenses, but on the ability of the user to use the proper objective, ocular, condenser, and illumination, and on controlling the aperture diaphragm and field diaphragms to their optimum settings for the system being used. It is not always possible to achieve the resolutions of a given lens even after proper manipulation because the detail is obscured by the background and the surrounding parts of the specimen. Technique in preparing the material being studied is of the utmost importance. Proper staining, thickness and mounting also determine how much we will resolve.

Resolution, angular aperture and wavelength of light being used are related. The angle of the cone of light is dependent on the refractive index of the medium between the objective front lens and the glass cover slip over the specimen (Fig. 6). Present-day objectives have clearly marked on their mounts the relationship between the angular cone and refractive index of the working medium, expressed in terms of numerical aperture (*N.A.*)

$$N.A. = i \sin \theta$$

where θ is one-half the angle of the cone of light entering the objective and i is the refractive index of the medium in which the objective is working. From this formula theoretically one

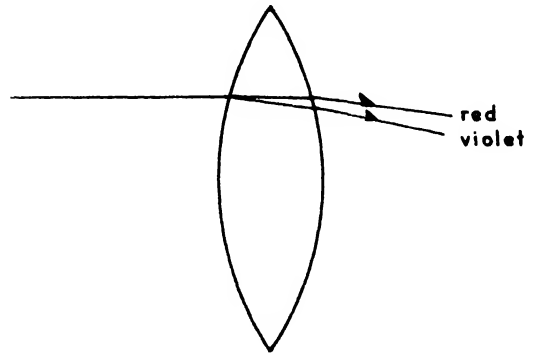


FIG. 3. Chromatic aberration.

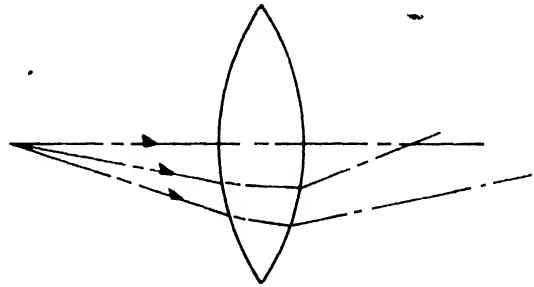


FIG. 4. Spherical aberration.

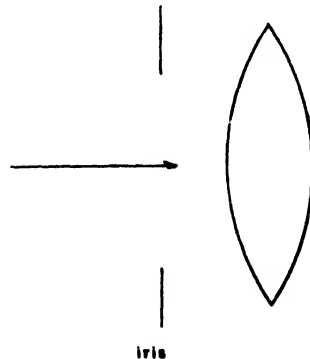


FIG. 5.

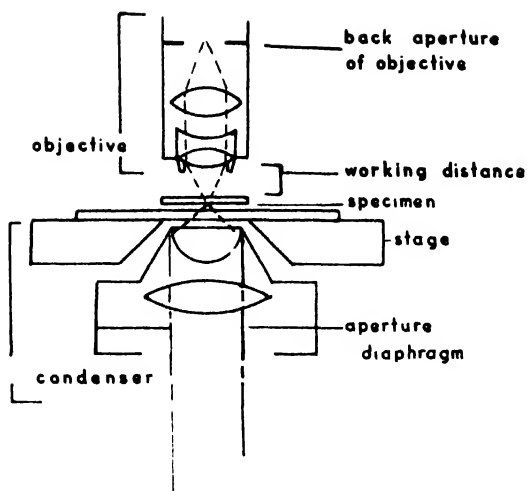


FIG. 6.

can see that the *N.A.* in air can never be greater than 1 because *i* would represent air with a refractive index of 1.

The relationship of *N.A.* and resolving power is as follows:

$$R = N.A. / \lambda$$

where λ is wavelength of light and *R* is the number of lines being separated by the lens.

There are three main classes of objectives: achromatic, fluorite, and apochromatic; the achromats are corrected for two colors (red and green) chromatically and one color spherically. To obtain more correction, substitution of the sandless glasses such as fluorite gives a little better correction than the achromats. The apochromatic objectives produce the best images, those finest in color correction and clarity. Apochromatic objectives are corrected for three colors chromatically and two spherically. To achieve the added value in the apochromatic objectives, they must utilize compensating eyepieces to correct for the color spread and a corrected condenser of a numerical aperture equal to or greater than the objective numerical aperture.

When using the microscope, it is very important to consider the working distance, depth of field and coverglass thickness. The working distance is the amount of free space from the front of the objective to the coverglass, and as the power and the *N.A.* increase, the working distance becomes shorter. Therefore, when a counting chamber with its normal coverglass is used, magnifications up to about 45X and *N.A.*'s up to about .66 can be used to view specimens. Objectives of higher magnification and larger *N.A.*'s reduce the working distance sufficiently to cause breakage of cover slips or inability to focus sharply on the specimen.

Depth of focus is the amount of thickness of the specimen that is in focus at one time. This amount decreases as magnification and *N.A.*

increase. In photomicrography this is very important and the specimen must be thin enough or parts of the specimen will be out of focus, obscuring detail of the portion that is in sharp focus. High-power dry objectives with correction collars to compensate for coverglass thickness should be used with cover glasses other than .18 mm to have optimum correction for spherical aberration. Objectives without the coverglass correction collar are designed and optically corrected for the use of coverglasses of .18 mm thickness. Coverglasses of other thickness increase the aberration, especially when using the high dry objectives.

Eyepieces. The eyepiece acts as the telescope of the system merely to remagnify the primary image of the objective; thus eyepiece magnification adds nothing to the resolution of the system. The practical rule is that one should not normally magnify more than 1000X the *N.A.* of the objective. More magnification is normally called empty magnification. However, in some instances to aid in counting, measuring, or drawing, ease of use can be achieved by using a higher-power eyepiece. Many companies are now making eyepieces with high eye relief so that persons requiring the additional distance between the eye lens and the Ramsden circle of the eyepiece can continue to wear their glasses while using the microscope.

There are many types of eyepieces. Four of the most common are Huygenian, periplane, widefield and compensating. The Huygenian is by far the simplest in design having 2 elements and being the least corrected, but most commonly, used eyepiece. Periplane eyepieces have greater correction thus producing better images as to color and flatness, while having about the same field of view as the Huygenian. Compensating eyepieces were designed for use with the apochromatic objectives to compensate for the objective deficiencies. In recent years the need to view more has brought about the development of eyepieces which cover greater fields of view.

Condensers. The condenser is commonly the most misused part of the microscope. A condenser has one optimum setting as to height and aperture diaphragm setting which controls the *N.A.* for each objective. The condenser height and aperture diaphragm should not be used to control the intensity of the illumination. If the aperture diaphragm is opened too far, the specimen is flooded with light and detail is washed out. If closed too far, the *N.A.* and detail are lost and diffraction becomes quite apparent.

Illuminators. The illuminator is a very essential part of the microscope. If one is to achieve the maximum from the microscope in terms of resolution, clarity and ease on the eyes of the user, a good illuminator must be employed. An illuminator must have a condensing system large enough to accommodate the apertures of the microscope and an iris diaphragm, sometimes referred to as a field stop. The condensing system and field stop must be focusable. To achieve the ultimate in illumination, filters and adjustable

mirrors in the illuminator help to adjust for maximum light through the system.

In the early days of microscopy, critical illumination was used by most microscopists. The source of light was broad such as the sun, reflection of the sun on white clouds or oil wick lamps. The source was focused so it would appear in the field of view (Fig. 7). Of course, the disadvantages of this system were size of the source

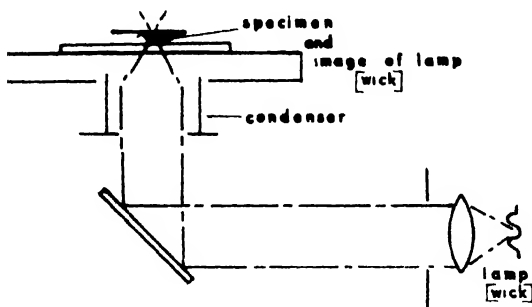


FIG. 7. Critical illumination.

and unevenness of illumination. The introduction of the concentrated coil filament lamps made use of critical illumination impractical due to the fact that the coil source was small and did not fill the aperture. Kohler solved this problem and the system is named after him. In Kohler illumination, the filament of the lamp is focused on the condenser iris (often called the aperture stop), and the field iris is focused so it will be in focus at the same plane as the specimen viewed through the microscope (Fig. 8).

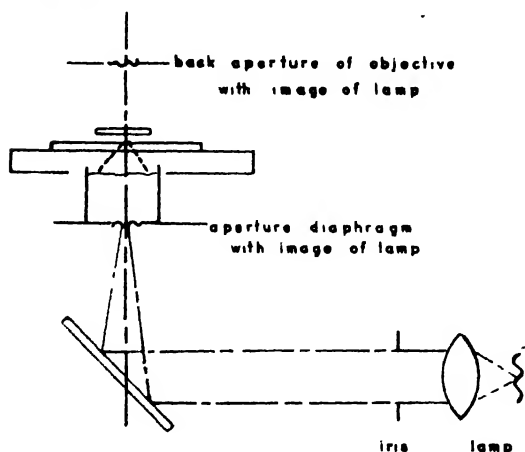


FIG. 8. Kohler illumination.

The use of filters makes it possible to change the visual appearance of the specimen. The use of a green filter for black and white photomicrography will, in many cases, increase the contrast and produce an excellent photograph.

The proper way to control the intensity of

illumination is to use neutral density filters. These filters control the intensity without changing the color balance of the source, e.g., if the source emits 20 per cent blue, 50 per cent green and 30 per cent red without a filter, then when a 50 per cent filter is placed in the system, only 50 per cent of the original light comes through the filter, but this 50 per cent is still composed of 20 per cent blue, 50 per cent green and 30 per cent red. We have cut down only the total intensity, not the balance.

Types of Microscope. There are many types of microscopes. Basically they are all similar, i.e., they have one or more objectives, eyepieces, condensers and some mechanical means to focus and manipulate the specimens. Often, in order to change from one type of microscope to another, only accessories have to be added.

The phase contrast method of microscopy was introduced in 1935 by Professor F. Zernike. Phase contrast is most useful in research and industry where an approach, other than the normal one, is necessary. The main advantage lies in the fact that contrast can be greatly enhanced in living and or unstained material, where, normally, little or no contrast exists with a bright field microscope. Direct observation of living material, such as yeast, bacteria and fungi, has made possible rapid, positive identification of specific organisms as well as observation of phenomena within the cells that have gone unnoticed for many years. Phase contrast relies on the combining of light waves and is dependent on amplitude and phase of these waves.

Dark field microscopy consists of a condenser with an opaque central stop which will not allow any direct light to enter the objectives. The circular cone of light is focused on the specimen which becomes very bright on a black background. Dark field shows objects that are too small to be seen by the bright field microscope, it is often used in the identification of spirochaetes, colloidal materials and chemical bonding.

Polarizing microscopes, in addition to having the normal components, have a polarizer and an analyzer. The light enters the polarizer where it is forced to vibrate in one direction. The light then passes through the condenser and objective to the analyzer which is set with the direction of vibration 90° to the polarizer; consequently no light is allowed to pass to the eyepiece. When material is placed in the system at the specimen plane, the material, in many cases, rotates the plane of polarization. Now that the polarized light has been rotated, a portion will pass through the eyepiece and be observed. Crystals that rotate the plane of polarization become either colored or white on a dark background. Crystals are classified by the way that they look under polarized light and by the amount of rotation which differs for each crystal. Other properties of crystals that can be studied by the polarizing microscopes are optical sign, extinction angle, birefringence, refractive index and pleochroism (see POLARIZED LIGHT).

The stereoscopic microscope is basically two microscopes, one for each eye. These two microscopes are mounted conveniently in a single housing for easy use. Each eye views the same specimen from a different angle producing a stereoscopic, three-dimensional view. The view through this type of microscope gives the user greater depth, erect image and larger fields than with the compound microscope. These instruments are used in dissection, genetic studies, and the assembly of small parts.

There are many universities, companies and individuals that have advanced the microscope to its present form. It is certainly one of the most important factors in our knowledge of bacteria and other medical sciences. Industry, in its quest for quality today, is investigating more closely the minute workings of its products, and with the advent of space exploration, the need for miniaturization has emphasized the part the microscope can play in studying, understanding and manipulating specimens.

DONALD A. BURGH

References

- Spitta, E. J., "Microscopy," Third edition, London, John Murray, 1924.
 Belling, J., "The Use of the Microscope," New York, McGraw-Hill Book Co., 1930.
 Gage, S. H., "The Microscope," Seventeenth edition, Ithaca, Comstock, 1943.
 Gray, P., "Handbook of Basic Microtechnique," Second edition, New York, McGraw-Hill Book Co., 1958.
 Allen, R. M., "Photomicrography," Second edition, New York, D. Van Nostrand, 1959.
 Bennett, Osterberg, Jupnik, and Richard, "Phase Microscopy," New York, John Wiley & Sons, 1951.

Cross-references: ABERRATION, DIFFRACTION THEORY OF, DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS, ELECTRON MICROSCOPE; LENS, OPTICAL INSTRUMENTS; REFRACTION.

MICROWAVE SPECTROSCOPY

Microwaves are electromagnetic waves which range in length from about 30 cm to a fraction of a millimeter or in frequency from 10^9 to 0.5×10^{12} cps. This corresponds to the rotational frequency range of a large class of molecules. Thus, microwave radiation passing through a gas can be absorbed when the rotating electric dipole moment of the molecule interacts with the electric vector of the radiation. Likewise, absorption can take place if the rotating magnetic moment of the molecule interacts with the magnetic vector of the radiation.

Most microwave spectroscopy is based on a study of transitions induced by interaction of the molecular electric dipole with the incident radiation.

A microwave spectrometer then consists basically of a monochromatic microwave source (klystron), an absorption cell, and a detector. The

absorption cell must transmit the microwave of interest and in the centimeter region may have cross-sectional dimensions of 1×4 cm and may be a few meters in length. Normally a metal strip is inserted along the length of the cell and is insulated from the cell. In this way, an auxiliary, spatially uniform electric field may be established in the absorption cell without affecting the microwaves. The Stark effect thereby produced splits the molecular energy levels into a series of levels and enables one to identify the transition.

The Hamiltonian for a rotating rigid asymmetric molecule including possible fine and hyperfine structure terms is given in Eq. (1). It is assumed, as is most commonly so, that the molecule is in a $^1\Sigma$ state, i.e., that there is no net electronic angular momentum and no net electron spin (singlet state). The Hamiltonian written is quite general and in many cases not all of the terms shown in Eq. (1) need be included to account for spectra observed under normal resolution.

A brief description will be given of each term in order that one may most simply understand the kinds of interactions which may occur and which may be pertinent to an understanding of the spectra of rotating molecules.

$$H = H_R + H_{\text{dist.}} + H_S + H_{\text{Ze}} + H_Q + H_{\text{Ze}} + H_D \quad (1)$$

1. H_R is the framework rotational kinetic energy and may be written

$$H_R = \frac{\hbar^2}{8\pi^2} \left[\frac{J_a^2}{I_a} + \frac{J_b^2}{I_b} + \frac{J_c^2}{I_c} \right]$$

where J_a , J_b , and J_c are the components of the total angular momentum in units of \hbar referred to body-fixed principal axes. I_a , I_b , and I_c are the moments of inertia about the respective principal axes. The Hamiltonian may be written

$$H_R = AJ_a^2 + BJ_b^2 + CJ_c^2$$

or displaying the total angular momentum J

$$H_R = AJ^2 + (B - A)J_b^2 + (C - A)J_c^2$$

A , B and C are rotational constants with $A > B > C$. In this form, units can be chosen so as to give energy levels directly in megacycles per second. H_R describes a rigid symmetric top if I_a is equal to I_b . For a diatomic molecule, $I_a = I_b$ and $I_c \gg 0$. ($1/I_c$ becomes very large, and the rotational levels about the c axis are too far apart to become excited by microwaves). For the spherical rotor, $I_a = I_b = I_c$, but this implies no dipole moment and therefore no observable rotational spectrum.

Energy levels for the symmetric top may be determined by analytical methods,¹ by factorization method², or by using the commutation properties of the angular momenta.³ In Eulerian coordinates, the wave equation separates, and the wave function has the form

$$\psi_J = N(JKM)e^{iK\phi}e^{iM\psi}\Theta_{JKM}(\theta)$$

Where Θ_{JKM} is the solution to the differential

equation in the polar angle θ which results after separation of the simple terms in the azimuthal angle ϕ and the angle ψ which defines the direction of the line of nodes. The equation for Θ_{JKM} with appropriate change of variable becomes the equation for the Jacobi polynomials.

The solution ψ is characterized by three quantum numbers: J , the total angular momentum; K the component of angular momentum along the symmetry axis of the molecule; and M , the magnetic quantum number or projection of J along an arbitrary space axis. The energy does not depend on M in the absence of external electric or magnetic fields. The energy levels for the symmetric top have the form

$$E_{J,K} = BJ(J+1) + (C-B)K^2$$

The rotational constants $A \approx B$ and C may typically range from 2000 to 300 000 Mc for presently observable spectra.

Selection Rules. In all but the accidentally symmetric top, the permanent dipole will be along the symmetry axis. In this case, the selection rules for absorption of radiation through rotation are:¹

$$J \rightarrow J \pm 1, K \rightarrow K$$

For a component of the dipole moment perpendicular to the symmetry axis, the selection rules are

$$\Delta J = \pm 1, 0 \text{ and } \Delta K = \pm 1$$

In both cases, $\Delta M = \pm 1, 0$.

The wave functions for the asymmetric top are expressed as linear combinations of symmetric top functions. The energy remains diagonal in J but not in K . One must, therefore, arbitrarily label the energy levels and determine the selection rules. This will not be done here. In order to do this, however, one needs to know only the nonvanishing matrix elements of the three components of the dipole moment for the symmetric top given above.

The selection rule for diatomic molecules is simply $J \rightarrow J \pm 1, M \rightarrow M, M \pm 1$.

2. $H_{\text{dist.}}$ describes centrifugal stretching corrections to the energy levels which for an asymmetric molecule can be quite complicated. Corrections for a symmetric top molecule are easily derived, and the framework energy in this case including centrifugal stretching is given by:

$$E_{J,K} = BJ(J+1) + (C-B)K^2 - D_J J^2(J+1)^2 - D_{JK} J(J+1)K^2 - D_K K^4$$

For a non-rigid diatomic molecule or linear polyatomic molecule, $K = 0$, and the energy is given by:

$$E_J = B_v J(J+1) - D_v J^2(J+1)^2$$

and since $J \rightarrow J+1$, for absorption the line frequencies are:

$$\nu_r = 2B_v(J+1) - 4D_v(J+1)^3$$

where B_v is the "effective" spectral constant,

$h/(8\pi^2 I_v)$, for the particular vibrational state for which the rotational spectrum is observed, and where D_v is the centrifugal stretching constant for that state.

In terms of the constants B_e and D_e for the hypothetical vibrationless state, B_v and D_v for diatomic molecules are:

$$B_v = B_e - \alpha(v + \frac{1}{2})$$

$$D_v = D_e + \beta(v + \frac{1}{2})$$

where α and β are interaction constants which are very small in comparison with B and D , respectively, and where v is the vibrational quantum number. For linear polyatomic molecules, there is more than one vibrational mode, and the above equations must be written in the more general form

$$B_v = B_e + \sum_i \alpha_i \left(v_i + \frac{d_i}{2} \right)$$

$$D_v = D_e + \sum_i \beta_i \left(v_i + \frac{d_i}{2} \right)$$

where the summation is taken over all the fundamental modes of vibrations. The subscript i refers to the i th mode, and d_i represents the degeneracy of that mode.

In the analysis of spectra, the variations of the stretching constants D_J and D_{JK} with vibrational state are customarily neglected. It is seldom possible to obtain sufficient data for the evaluation of these effects upon B and for determination of B_e even for the simpler symmetric tops.

Centrifugal stretching constants may typically range from 8.5 to 0.002 Mc or less.⁵

3. The third term H_S is the contribution to the Hamiltonian arising from the Stark effect and may be written as

$$H_S = \mu_e \cdot \mathcal{E}$$

where μ_e is the vector dipole moment and \mathcal{E} is the external electric field.

If the dipole moment lies along the "c" body-fixed principal axis, H_S has the form

$$H_S = \mu_e (A_x^c \mathcal{E}_x + A_y^c \mathcal{E}_y + A_z^c \mathcal{E}_z)$$

where A_x^c, A_y^c, A_z^c are the direction cosines of the c principal axis with space-fixed axes xyz. $\mathcal{E}_x, \mathcal{E}_y$, and \mathcal{E}_z are the components of the electric field along the space-fixed axes. Additional terms will be added to this expression if the dipole moment has components along the remaining two principal axes. In order to obtain the contribution to the energy from this part of the Hamiltonian, one must evaluate matrix elements of the direction cosines with respect to symmetric top wave functions. Methods described in reference 3 enable one to do this.

In the case of a symmetric top, the dipole moment will have a component only along the c axis so that H_S consists of the single term $A_x^c \mathcal{E}_x$ when $\mathcal{E}_x = \mathcal{E}_y = 0$.

In this case, the energy associated with H_S is

diagonal in all three quantum numbers JKM and has the form

$$E_s = -\mathcal{E}_z \mu_e \frac{MK}{J(J+1)}$$

where M , the "magnetic quantum number," measures the component of J along \mathcal{E}_z and can take the values $M = J, J-1, \dots, -J$. The selection rules for M are $M \rightarrow M$; $M \rightarrow M \pm 1$, depending on the polarization of the microwaves.

For asymmetric molecules, μ_e will in general have components along the A and B axes as well as C . The A and B components give rise to matrix elements off diagonal in J , K and M . For a dipole moment of 1 debye and an electric field of 1 volt/cm, $\mu_e \mathcal{E}$ is 0.5 Mc.

4. H_{ze} is the contribution to the Hamiltonian due to the interaction of the external magnetic field with the magnetic moment which is created by rotation of the molecule.

We also include the interaction of the external magnetic field with the dipole moment of individual nuclei. For a molecule with two nuclear spins, I_1 and I_2 , this may be written

$$H_{zi} = \sum_{j,k} \mu_n(J)_{jk} J_{jk} + \mu_n g_1 (I_1 \cdot H) + \mu_n g_2 (I_2 \cdot H)$$

where $g(J)$ is in general a tensor, J_j are components of J along axes to which H is referred, H_k are the components of the field H usually referred to the space-fixed axes, and g_1 and g_2 are called the nuclear magnetic g factors. The interaction between J and H is the same order as that between the nuclear spin and H . Therefore, we introduce the term μ_n so that g coefficients are of the order of unity. Thus for a field of one gauss, the quantity $\mu_n g H$ is 0.7 kc.

For molecules with electronic angular momentum, μ_n in the first term is replaced by μ_0 , the Bohr magneton which is 1836 times larger than μ_n . Thus, in this case $\mu_n g H$ is ~ 1.4 Mc for a field of one gauss.

5. H_Q is the energy of interaction of the nuclear electric quadrupole with the gradient of the electric field produced by the electrons in the molecule at nucleus with spin I . For a nucleus on the axis of a symmetric top, the quadrupole operator is ordinarily considered to be of the form:

$$H_Q = \frac{-eQq}{2J(2I-1)(2J-1)(2J+3)} \left\{ 1 - \frac{3K^2}{J(J+1)} \right\} \left[3(J+1)^2 + \frac{3}{2}(J \cdot I) - J^2 I^2 \right]$$

This operator yields only those matrix elements of the quadrupole interaction which are diagonal in J . The diagonal contributions are sufficient for most cases.

In the expression above, eQ is defined by

$$eQ = (I, I | \int \rho_n [3z_n^2 - r_n^2] d\tau_n | I, I)$$

where ρ_n is the nuclear charge density at a distance r_n from the center of charge of the nucleus and $d\tau$ is the differential volume element for the

nuclear volume. z_n is the position coordinate along the direction of the nuclear spin I . The matrix element considered is that for which $M_I = I$.

The quantity q is defined as

$$q = \left[\frac{\partial^2 V}{\partial c^2} \right]_{(r_n \rightarrow 0)}$$

where V is the electrostatic potential due to the electronic cloud and other nuclei surrounding the nucleus and c is the axis in the body-fixed system which is parallel to the symmetry axis of the molecule. The quantity eqQ varies from ~ 1000 to 1000Mc/sec although the intermediate values are more common.⁵

6. H_{zi} represents the interaction between the magnetic field caused by rotation of the charged particles which make up the molecule and the nuclear magnetic moments of the nuclei. For the case of 2 nuclei, this takes the form

$$H_{zi} = \sum_{j,k} C(1)_{jk} J_{jk} I_{1k} + \sum_{j,k} C(2)_{jk} J_{jk} I_{2k}$$

$C(1)$ and $C(2)$ represent the internal magnetic moment tensors for the two nuclei. The C coefficients are of the order of 10^{-2}Mc^2 . J and I are pure numbers. This correction will therefore be unimportant for the large majority of molecules.

7. H_D is the dipole interaction between the two nuclei which may be written in the form

$$H_D = \frac{g_1 g_2 \mu_n^2}{R^3} \left[I_1 \cdot I_2 - \frac{3(I_1 \cdot R)(I_2 \cdot R)}{R^2} \right]$$

where g_1 and g_2 are the nuclear gyromagnetic ratios of the nuclei, μ_n is the nuclear magneton and R is the distance between the two nuclei.

The operator which is usually used to represent this interaction is

$$H_D = \frac{g_1 g_2 \mu_n^2}{R^3} \frac{[3(I_1 \cdot J)(I_2 \cdot J) + 3(I_2 \cdot J)(I \cdot J - 2)I_1 \cdot I_2 J^2]}{(2J-1)(2J+3)}$$

This operator, like that given for the quadrupole interaction above, and usually quoted in the literature, will yield only those matrix elements which are diagonal in the quantum number J . The coefficient $g_1 g_2 \mu_n^2 / R^3$ may be of the order of a kilocycle. This correction is observed only in very rare cases.⁷

Matrix elements for all of the above-mentioned components of the Hamiltonian may be evaluated by the methods in references 6. The matrix elements themselves are too lengthy to be tabulated here.

There is an additional interaction which, for completeness should be mentioned. The nuclear spins may interact with one another through mutual coupling with the surrounding electron cloud. This gives a correction of the form $C I_1 \cdot I_2$. The coefficient C may be larger than that in the dipole-dipole interaction term.

The preceding discussion emphasizes the measurement and interpretation of microwave absorption spectra. One is thereby led to a knowledge of the structure of the molecule, the value of nuclear spins, and various coupling constants.

Microwave spectroscopy is also an effective technique for determination of barrier heights associated with internal rotation.^{8,9,10}

A discussion of line shapes and intensities has been omitted. We have not discussed electronic magnetic resonance, that is, absorption associated with a net nonzero electron spin. Absorptions of this type are observed in molecules, free atoms, radicals, and solids.

For further details and a discussion of topics omitted, the reader is referred to the excellent article of W. Gordy⁸ and the text⁵ by Townes and Schalow on microwave spectroscopy.

DONALD G. BURKHARD

References

1. Dennison, D. M., *Phys. Rev.*, **28**, 318 (1926); Reiche, F., and Rademacher, H., *Z. Physik*, **39**, 444 (1926) and **41**, 453 (1927).
2. Burkhard, D. G., *J. Mol. Spectry.*, **2**, 187 (1958); Shaffer, W. H., and Louck, J. D., *J. Mol. Spectry.*, **3**, 123 (1959).
3. Klein, O., *Z. Physik*, **58**, 730 (1929).
4. Dennison, D. M., *Rev. Mod. Phys.*, **3**, 280 (1931).
5. Townes, C. H., and Schawlow, A. L., "Microwave Spectroscopy," New York, McGraw-Hill Book Co., 1955.
6. Condon, E. U., and Shortley, G. H., "Theory of Atomic Spectra," Cambridge, England, The University Press, 1935, reprinted with corrections 1951.
7. Landau, L. D., and Lifshitz, E. M., "Quantum Mechanics," Reading, Mass., Addison and Wesley 1958.
8. Thaddeus, P., Krisher, L. C., and Loubser, J. M. N., *J. Chem. Phys.*, **40**, 257 (1964).
9. Gordy, Walter, "Microwave Spectroscopy," in "Handbuch der Physik," Vol. XXVIII, Berlin, Springer-Verlag, 1957.
10. Lin, C. C., and Swalen, J. D., *Rev. Mod. Phys.*, **31**, 841 (1959).
11. Burkhard, D. G., *J. Opt. Soc. Am.*, **50**, 1214 (1960).

Cross-references: SPECTROSCOPY, ZEEMAN AND STARK EFFECTS.

MICROWAVE TRANSMISSION

That portion of the electromagnetic spectrum which lies adjacent to the far-infrared region is commonly identified by the term microwaves. The shortest and longest practical microwave wavelengths are in the vicinity of 1 mm (3×10^{11} cps or 3×10^5 mc/sec, abbreviated 300 000 Mc or 300 gc) and 10 cm (3×10^9 cps, abbreviated 3000 Mc or 3 Gc), respectively.

The development of microwave transmission on a large scale began in 1940 with the advent of the magnetron, an electronic generator of high-power microwaves. The magnetron made possible

wartime radar at approximately 3000 Mc and led to the utilization of waveguides for the efficient transmission of microwaves from the generator to the transmitting antenna and from the receiving antenna to the detector.

In essence, a waveguide is a hollow metal tube capable of propagating electromagnetic waves within its interior from its sending end to its receiving end. Unlike waves in space which propagate outward in all directions, waves in waveguides are fully confined while they propagate. By far, the most commonly used waveguide is rectangular in cross section. Such rectangular waveguides have the property of maintaining wave polarization over wide frequency ranges.

Propagation in a waveguide is limited by the cross-section dimensions of the guide. In a rectangular guide, the highest wavelength that can be propagated is equivalent to twice its width. For example, to transmit a 3000-Mc signal in a rectangular waveguide, the width of the guide must be at least 5 cm. At 300 Mc, this width becomes 50 cm. Thus, for the transmission of microwaves, waveguides are reasonable in size. Furthermore, when used with ferrites and magnets, waveguide components can be designed to function as isolators, circulators, modulators, discriminators, or attenuators.

The propagation of radio signals in space between transmitting and receiving antennas can be described in terms of ground waves, sky waves and space waves. At microwave frequencies, ground waves attenuate completely within a few feet of travel, sky waves are influenced by the ionosphere and can penetrate through into outer space, and space waves behave like light waves as they travel through the atmosphere immediately above the surface of the earth. Microwave space waves travel in a direct line of sight, can be reflected from smooth conducting surfaces, and can be focused by reflectors or lenses. Their behavior is similar to light waves and they follow many of the rules of optics.

If a space wave is radiated from a point antenna, the radiated energy spreads out like an ever-expanding sphere, and the amount of energy per square foot of wave front decreases inversely with the square of the distance from the antenna. The power that can be extracted from a wave front by a similar point antenna varies inversely with the square of the frequency. Thus, a point antenna receives power which is inversely proportional to both the square of the distance from the source and the square of the frequency. The ratio of the power received to the total power radiated is known as path attenuation.

When the receiving antenna is a parabola-shaped dish, the power extracted from the wave front is greatly increased. The ratio of the power received by such an antenna to the power received by a theoretical point antenna is defined as antenna gain. The gain of a parabolic antenna increases with the antenna area and the operating frequency. Thus, for a given microwave path with fixed-size antennas, the path attenuation increases with frequency, the antenna gain increases with

frequency, and the over-all result is that one tends to offset the other.

In broadcast radio, signal power radiates equally in all directions and a receiving antenna picks up only a tiny fraction of the signal power. To overcome this low efficiency, the broadcast station must transmit a large amount of power. By contrast, a point-to-point microwave system radiates only a small amount of power, but it uses a directional transmitting antenna to concentrate power into a narrow beam directed toward the receiving antenna. Consequently, such systems are characterized by high efficiencies.

Because microwave transmission follows essentially a straight line, the distance between terminals tends to be limited by the curvature of the earth and by obstructions such as trees, buildings and mountains. Reflectors are often used to redirect a beam over or around an obstruction. The simplest and most common reflector system consists of a parabolic antenna mounted at ground level which focuses a beam on a reflector mounted at the top of a tower. This reflector inclined at 45° redirects the beam horizontally to a distant site where a similar "periscope" reflector system may be used to reflect the beam down to another ground level.

Often a tower cannot provide clearance over an obstruction. For example, if two sites are separated by a mountain, it may be necessary to use a large, flat surface reflector referred to as a "billboard" reflector. In a typical system, a "billboard" reflector might be located at a turn in a valley, effectively bending the beam to follow the valley. Many arrangements are possible which, in effect, resemble huge mirror systems.

Microwaves are ideally suited for communication systems where a broad frequency bandwidth of the order of several megacycles is required for the rapid transmission of signals which contain a large amount of information, such as in television. Most of the major cities of the United States are serviced by microwave television links so that they can receive television programs which originate from other cities. These systems can also accommodate thousands of telephone channels.

In 1960, experiments were initiated which aimed toward communicating over transoceanic distances via microwaves by utilizing orbiting balloons as reflectors. Echo I and Echo II were attempts in this direction.

Microwaves are broadly used for radar, navigation, and for the launching, guidance, and fusing of missiles. A typical defense project which uses microwave techniques is the DEW radar line which protects the United States from external enemy attacks.

Of recent vintage, is the U.S. Air Force's HAYSTACK facility being built at Millstone Hill in Massachusetts for spacecraft tracking, space communications, and radar astronomy. This facility is expected to track a target the size of a dime at a distance of 1000 miles. It will be the first western radar capable of making contact with more distant planets, such as Mars, Mercury

and Jupiter. It will operate at a microwave frequency of 27 500 Mc.

ANTHONY B. GIORDANO

References

- Nichols, E. J., and Tear, J. D., "Joining the Infrared and Electric Wave Spectra," *Astrophys. J.*, **61**, 17-37 (1923).
Carter, S. P., and Solomon, L., "Modern Microwaves," *Electronics* (June 24, 1960).
Southworth, G. C., "Survey and History of the Progress of the Microwave Art," *Proc. IRE* (May 1962).

Cross-references: ANTENNAS, ELECTROMAGNETIC THEORY, MICROWAVE SPECTROSCOPY, PROPAGATION OF ELECTROMAGNETIC WAVES, RADAR, WAVEGUIDES.

MODULATION

Modulation is defined as the process, or the result of the process, whereby some characteristic of one wave is varied in accordance with some characteristic of another wave (ASA). Usually one of these waves is considered to be a carrier wave while the other is a modulating signal. The various types of modulation, such as amplitude, frequency, phase, pulse width, pulse time, and so on are designated in accordance with the parameter of the carrier which is being varied.

Amplitude modulation (AM) is easily accomplished and widely used. Inspection of Fig. 1 shows that the voltage of the amplitude modulated wave may be expressed by the following equation

$$v = V_c(1 + M \sin \omega_m t) \sin \omega_c t$$

where ω_c and ω_m are the radian frequencies of the carrier and modulating signals, respectively. The modulation index M may have values from zero to one. When the trigonometric identity $\sin a \sin b = \frac{1}{2} \cos(a-b) - \frac{1}{2} \cos(a+b)$ is used in the equation above, this equation becomes

$$v = V_c \sin \omega_c t + \frac{MV_c}{2} \cos(\omega_c - \omega_m)t - \frac{MV_c}{2} \cos(\omega_c + \omega_m)t$$

This equation shows that new frequencies, called side frequencies or side bands, are generated by the amplitude modulation process. These new frequencies are the sum and difference of the carrier and modulating frequencies.

Amplitude modulation is accomplished by mixing the carrier and modulating signals in a nonlinear device such as a vacuum tube or transistor amplifier operated in a nonlinear region of its characteristics. The nonlinear characteristic produces the new side-band frequencies. Frequency converters or translators and AM detectors are basically modulators. The various types of pulse modulation are actually special types of amplitude modulation.

Frequency modulation (FM) is illustrated by Fig. 2. The frequency variation, or deviation, is

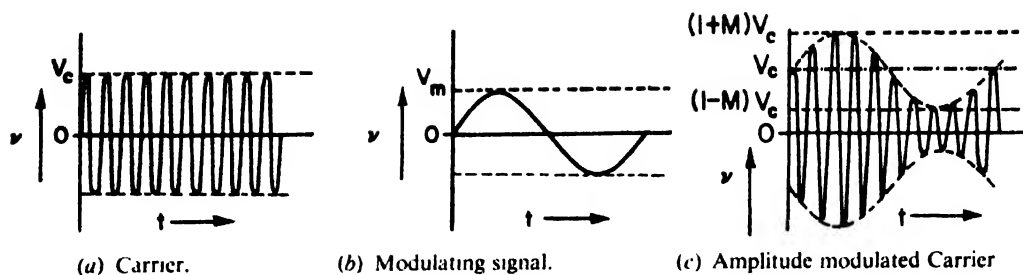


FIG. 1. Amplitude modulation.

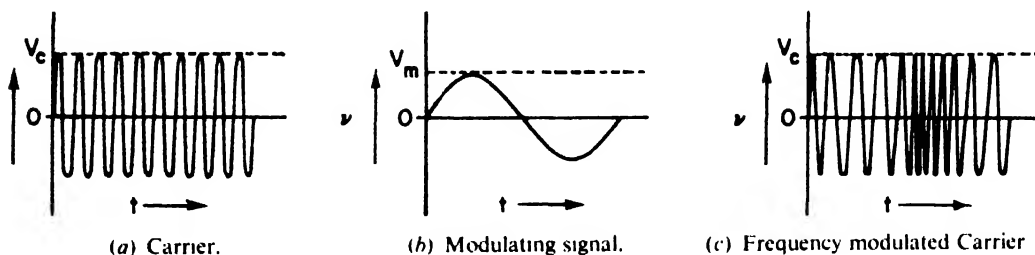


FIG. 2. Frequency modulation.

proportional to the amplitude of the modulating signal. The voltage equation for a frequency modulated wave follows.

$$v = V_c \sin(\omega_c t + M_f \sin \omega_m t)$$

The modulation index M_f is the ratio of maximum carrier frequency deviation to the modulating frequency. This ratio is known as the deviation ratio and may vary from zero to values of the order of 1000. FM requires a broader transmission bandwidth than AM but may have superior noise and interference rejection capabilities. A large value of modulation index provides excellent interference rejection capability but requires a comparatively large bandwidth. The approximate bandwidth requirement for a frequency modulated wave may be obtained from the following relationship

$$\text{Bandwidth} = 2 (\text{Modulating frequency}) (M_f + 1)$$

The noise and interference characteristics of FM transmission are normally considered satisfactory when the modulation index or deviation ratio is five or greater.

Phase modulation is accomplished when the relative phase of the carrier is varied in accordance with the amplitude of the modulating signal. Since frequency is the time rate of change of phase, frequency modulation occurs when the phase modulating technique is used and vice versa. In fact, the equation given for a frequency-modulated wave is equally applicable for a phase-modulated wave. However, the phase-modulating technique results in a deviation ratio, or modulation index, which is independent of the modulating frequency, while the frequency modulating technique results in a deviation ratio which is inversely

proportional to the modulating frequency, assuming invariant modulating voltage amplitude in each case.

The phase-modulating technique can be used to produce frequency-modulated waves, providing the amplitude of the modulating voltage is inversely proportional to the modulating frequency. This inverse relationship can be obtained by including, in the modulator, a circuit which has a voltage transfer ratio inversely proportional to the frequency.

CHARLES L. ALLEY

References

- Alley, C. L., and Atwood, K. W., "Electronic Engineering," New York, John Wiley & Sons, 1962.
- Terman, F. E., "Radio Engineers Handbook," New York, McGraw-Hill Book Co., 1943.
- Hund, August, "Frequency Modulation," New York, McGraw-Hill Book Co., 1942.
- Lurch, E. N., "Fundamentals of Electronics," New York, John Wiley & Sons, 1960.

MOLE CONCEPT

The mole (derived from the Latin *mole* - heap or pile) is the chemist's measure of amount of pure substance. It is relevant to recognize that the familiar *molecule* is a diminutive (little mole). Formerly, the connotation of *mole* was a "gram molecular weight." Current usage tends more to use the term *mole* to mean an amount containing Avogadro's number of whatever units are being considered. Thus, we can have a mole of atoms, ions, radicals, electrons or quanta. This usage makes unnecessary such terms as "gram-atom," "gram-formula weight," etc.

A definition of the term is: *The mole is the amount of (pure) substance containing the same number of chemical units as there are atoms in exactly twelve grams of C^{12} .* This definition involves the acceptance of two dictates—the scale of atomic masses and the magnitude of the gram. Both have been established by international agreement. Usage sometimes indicates a different mass unit, e.g., a “pound mole” or even a “ton mole”; substitution of “pound” or “ton” for “gram” in the above definition is implied.

All stoichiometry essentially is based on the evaluation of the number of moles of substance. The most common involves the measurement of mass. Thus 25.000 grams of H_2O will contain 25.000/18.015 moles of H_2O ; 25.000 grams of sodium will contain 25.000/22.990 moles of Na (atomic and formula weights used to five significant figures). The convenient measurements on gases are pressure, volume and temperature. Use of the ideal gas law constant R allows direct calculation of the number of moles $n = (P \cdot V)/(R \cdot T)$. T is the absolute temperature; R must be chosen in units appropriate for P , V and T (e.g. $R = 0.0820$ liter atm mole $^{-1}$ deg K^{-1}). It may be noted that acceptance of Avogadro's principle (equal volumes of gases under identical conditions contain equal numbers of molecules) is inherent in this calculation. So too are the approximations of the ideal gas law. Refined calculations can be made by using more correct equations of state.

Many chemical reactions are most conveniently carried out or measured in solution (e.g., by titration). The usual concentration convention is the *molar* solution. (Some chemists prefer to use the equivalent term *formal*). A 1.0 molar solution is one which contains one mole of solute per liter of solution. Thus the number of moles of solute in a sample will be

$$n = \text{Volume (liters)} \cdot \text{Molarity (moles liter)}$$

The amount of chemical reaction occurring at an electrode during an electrolysis can be expressed in moles simply as $n = q/(z \cdot \mathcal{F})$ where z is the oxidation number (charge) of the ion and \mathcal{F} is the faraday constant, 96 487.0 coulombs/mole. Thus the *faraday* can be considered to be the charge on a mole of electrons. This affords one of the most accurate methods of evaluating the avogadro number ($6.02252 \cdot 10^{23}$), since the value of the elementary charge is known with high precision.

Modern chemistry increasingly uses data at the atomic level for calculation at the molar level. Since the former often are expressed as quanta, appropriate conversion factors must involve the avogadro number. Thus the *einstein* of energy is that associated with a mole of quanta, or $E = Nh\nu$. Thus light of 2537 Å wavelength will represent energy of

Another convenient conversion factor is 1 eV/particle = 23.05 kcal/mole.

The chemist's use of formulas and equations always implies reactions of moles of material, thus $HCl(g)$ stands for one mole of hydrogen chloride in the gaseous state. Thermodynamic quantities are symbolized by capital letters standing for molar quantities, e.g., C_p (heat capacity at constant volume in cal mole $^{-1}$ deg $^{-1}$), G (Gibbs function in cal/mole), etc. At times it is more convenient to convert an extensive property into an intensive expression. This is especially true in dealing with multicomponent systems. These are referred to as “partial molal quantities” and are given a symbol employing a bar over the letter. Thus the partial molal volume, \bar{V}_1 ($\partial V/\partial n_1$) is the rate of change of the total volume of a solution with the amount (number of moles) of component 1.

WILLIAM F. KIEFFER

References

- Kieffer, W. F., “The Mole Concept in Chemistry,” New York, Reinhold Publishing Corp., 1962.
Lewis, G. N., and Randall, M., “Thermodynamics,” Second edition, revised by Pitzer, K. S., and Brewer, L., New York, McGraw-Hill Book Co., 1961.

Cross-references: CHEMISTRY, ELECTROCHEMISTRY, GAS LAWS, MOLECULAR WEIGHT.

MOLECULAR WEIGHT

The molecular weight of a chemical compound is the sum of the atomic weights of its constituent atoms. The molecule is the smallest weight of a substance which still retains all of its chemical properties. By convention, each atomic weight, and therefore molecular weights, are expressed relative to the arbitrary value, 16, for the oxygen atom. For example, the molecule of acetic acid, CH_3COOH , contains two atoms of carbon, four of hydrogen, and two of oxygen, so that its molecular weight is the sum of $2(12.01) + 4(1.01) + 2(16.00)$, which totals 60.06. This molecular weight value is clearly in arbitrary units, but a related quantity, the gram-molecular weight or mole, is the molecular weight expressed in grams. One mole of any compound has been found to contain $6.023 \cdot 10^{23}$ molecules, and this number is called the avogadro constant.

The most commonly used system of molecular weights, called “chemical,” is based on the value 16.00 for the naturally occurring mixture of isotopes of oxygen. For purposes where single molecules are to be compared, another system of “physical” molecular weights is sometimes used. This is based on the value 16.00 for

$$E = \frac{6.023 \cdot 10^{23}(\text{quanta/mole}) \cdot 6.62 \cdot 10^{-27}(\text{erg-sec}) \cdot 3.000 \cdot 10^{10}(\text{sec/cm})}{2.537 \cdot 10^{-5}(\text{cm}) \cdot 4.184 \cdot 10^7(\text{erg/cal}) \cdot 10^3(\text{cal/kcal})}$$

$$E = 113\text{kcal/mole}$$

the most abundant (99.8 per cent) oxygen isotope. Under this system, molecular weight values are about 1.0002 times those commonly used. A recent conference of the International Commission on Atomic Weight (1961) adopted a value of 12 for the carbon-12 isotope as a standard.

The weights of molecules range from a value of about two for the hydrogen molecule to several millions for some virus molecules and certain polymeric compounds. Molecular dimensions accordingly range from a diameter of about 4Å for the hydrogen molecule to several thousand angstroms—which has permitted viewing single large molecules in the electron microscope. Molecular sizes are generally much smaller and are not measured directly, but are deduced from x-ray diffraction studies of ordered groups of molecules in the crystalline state or from the physical properties such as hydrodynamic behavior of molecules in the gaseous or liquid state.

Many methods for determining molecular weights which are described below depend fundamentally on counting the number of molecules present in a given weight of sample. However, any usable sample contains a very large number of molecules: at least ten trillion of the largest known molecules are present in the smallest weight measurable on a sensitive balance. Therefore, an indirect count is made by measuring physical properties which are proportional to the large number of molecules present. A consequence of the large number of molecules sampled is the averaging of any variations in content of atomic isotopes in individual molecules, so that normal isotopic fluctuations lead to no measurable deviation of molecular weight values. Abnormally high concentrations of isotopes in radiation products may, however, produce altered molecular weights.

The term molecular weight is properly applied to compounds in which chemical bonding of all atoms holds the molecule together under normal conditions (see BOND, CHEMICAL). Thus, covalent compounds, as represented by many organic substances, usually are found to have the same molecular weight in the solid, liquid, and gaseous states. However, substances in which some bonds are highly polar may exist as unionized or even associated molecules in the gaseous state and in nonpolar solvents, but they may be ionized when dissolved in polar solvents. For example, ferric chloride exists in the gaseous state as FeCl_3 at high temperatures, as Fe_2Cl_6 at lower temperatures as well as in nonpolar solvents, but reverts to FeCl_3 in solvents of moderate polarity, and becomes ionic in water solutions—as chloride ions and hydrated ferric ion. Similarly, acetic acid and some other carboxylic acids associate as dimers in the vapor state and in solvents of low polarity, but exists as monomers with progressive ionization as the solvent polarity increases.

Truly ionic compounds, such as most salts, exist only as ions in the solid and dissolved states, so that the term molecule is not applicable and is not commonly used. Instead, the term,

formula weight, is used; this denotes the sum of the atomic weights in the simplest formula representation of the compound. If a broad definition of a molecule as an aggregate of atoms held together by primary valence bonds is adopted, then salts in the crystalline state would appear to have a molecular weight which is essentially infinite and limited only by the size of the crystal, since each ion is surrounded by several ions of opposite polarity to which it is attached by ionic bonds of equal magnitude.

A further complication in the definition of molecular weights occurs with inorganic polymers whose polymeric nature is clearly evident in both their crystal structure and their highly viscous behavior in the molten state. However, the magnitude of their molecular weights often cannot be found by conventional methods because they are either insoluble or react with solvents, with consequent degradation. These examples indicate that the molecular weight often depends on the conditions used for measurement and must be specified where compounds subject to association, dissociation or reaction are studied.

The history of the clarification of molecular weight concepts is of considerable interest, since this was so intimately related to other developments in chemical knowledge. Although Dalton had published a table of atomic weights in 1808, and by 1825 molecular formulas, derived from combining weights, were in use, many misconceptions of these formulas remained until about 1860. Then evidence from chemical reactions and from measurements of vapor densities firmly established the formulas of many inorganic and simple organic compounds as they are represented today. The vapor density method, based on Avogadro's hypothesis, was thus the first molecular weight method and continues to be useful for compounds that can be easily volatilized. It was not until 1881 that Raoult showed that the depression of freezing points was proportional to the molar concentration of solute. In 1884, van't Hoff related the osmotic pressure of solutions to the vapor pressure, boiling point, and freezing point behavior, and these methods were quickly put into use for determining molecular weights. The abnormal physical properties of salt solutions were explained in 1887 by the ionization theory of Arrhenius, and the very careful measurements of many of these properties furnished the strongest confirmation of the theory. While these measurements provided the most precise determinations of the extent of dissociation of weak electrolytes, they also contributed to the development of the Debye-Hückel theory for strong electrolytes.

Molecular Weight Distributions. There are many systems, particularly among the polymers and proteins, which consist of molecules of various chain lengths, and thus of various molecular weights—so-called polydisperse systems. In such cases, molecular weight values have an ambiguous meaning, and no single such value will completely represent a sample. Various techniques for measuring molecular weights, when applied to one

of these materials, will produce values which often disagree by a factor of two or more. This disagreement arises from the different bases of the methods—for example, some methods yield so-called number-average molecular weights by determining the number-concentration of molecules in a sample, while other methods produce weight-average molecular weights which are related to the weight-concentrations of each species. Another common value is the viscosity-average molecular weight, which is related to the viscosity contribution of each species. Other bases are of importance for certain methods of study, and some of these are complex functions involving several averages. For some purposes, the determination of a single average molecular weight is sufficient for establishing relations between molecular weight and the behavior of polymers, but the type of molecular weight average must be so chosen as to have a close relation to the behavior property of interest. A more detailed knowledge of the constitution of a sample is sometimes required, particularly if several properties are to be considered, or if unusual forms of molecular weight distribution curve are present.

The problem of completely defining the molecular weight nature of polydisperse materials is most accurately solved by determining the frequency of occurrence of each molecular species and representing the results as a frequency distribution curve. Such a study is generally quite tedious, though there are a few methods which provide much of the required information in one experiment. The most common method of approximating a molecular weight distribution curve is to separate the sample into a series of fractions which differ in chain length and then to determine the average molecular weight of each fraction. The distribution curve which can be drawn from these results is often inexact, but it is the best method which is commonly available. Obviously, such a study requires so much effort that it can be applied only to materials of great interest. Some recent improvements in fractionation methods based on the fractional solution of material finely dispersed in a column appear to offer greater speed of separation and improved sharpness of the fractions, which should overcome some objections. The ultracentrifuge is less commonly used for determining molecular weight distributions in a single experiment, partly because of its cost and the complexity involved in the analysis of data.

Uses. Molecular weight measurements, in conjunction with the law of combining proportions, have enabled the atomic weights of elements in compounds to be determined. When the atomic weights are known, molecular weight measurements permit the assignment of molecular formulas. Other applications to compounds of low molecular weight allow determination of the extent of ionization of weak electrolytes, and the extent of association of some uncharged compounds which aggregate. The study of molecular weights is becoming increasingly valuable in assessing the effects which various molecular

species of a polymer sample have on the physical properties of the product. Through such knowledge, the synthetic process may be modified to improve the properties of polymers.

Methods of Measurement. Many physical and certain chemical properties vary substantially with the molecular weight of compounds, and these properties are the bases of all molecular weight methods. The summary given in this section includes principally the methods which are most frequently used or have general applicability. The choice of the most suitable method for a given sample depends on its state (gas, liquid, or solid), the magnitude of the molecular weight and the accuracy required in its determination, as well as on the stability of the compound to physical or chemical treatment. Some mention of the applicability of the methods in these regards is given wherever possible.

Gases and Liquids. Avogadro's hypothesis (1811) that equal volumes of different gases contain the same number of molecules under the same conditions made it possible to find how many times heavier a single molecule of one gas is than that of another. Thus, relative molecular weights of all gases could be established by comparing the weights of equal volumes of gases. The significance of the idea and utilization of this method were first clearly demonstrated by Cannizzaro in 1858, and this represents the first available method for determining molecular weights. With the additional information from chemical experiments on the number of atoms of each kind present in each molecule, the relative weights of each atom were obtained. The assumption of the integral value, 16, for the atomic weight of oxygen (to give a value close to unity for the lightest element, hydrogen) then enabled molecular weights of all gaseous compounds to be determined. The method obviously can be applied to other molecules which normally occur in the liquid state but can be volatilized by heating. The Dumas and Victor Meyer methods are most used for molecular weight determinations with liquids in this way. These methods have been refined so that gas densities can now be determined with an accuracy of 0.02 per cent, and extremely small weights of material (about 1 μ g) can be similarly studied with somewhat less accuracy. High temperatures up to 2000°C have been used to study substances which are volatilized only with difficulty, provided decomposition can be avoided.

Solids Measured by Colligative Methods. It has been shown that nonvolatile molecules dissolved in a solvent affect several physical properties of the solvent in proportion to the number of solute molecules present per unit volume. Among these properties are a decrease of the vapor pressure of the solvent, a rise in its boiling point, a decrease in its freezing point, and the development of osmotic pressure when the solution is separated from the solvent by a semipermeable membrane. Properties such as these which are related to the number of molecules in a sample rather than to the type of molecule are called

colligative properties. They are the basis for some of the most useful techniques for molecular weight determination. The magnitude of the effects and the ease of measurement differ greatly, so that certain of the colligative properties are preferred for this purpose. For example, an aqueous solution containing 0.2 gram of sucrose (molecular weight 342) in 100 ml has a vapor pressure 0.01 per cent less than that of the solvent, a boiling point 0.003°C greater, and a freezing point 0.011°C lower than the solvent, but will develop an osmotic pressure of 150 cm of water. Since the effects are related to the number-concentration of solute molecules, each method leads to a number average molecular weight if the sample consists of a mixture of molecules of different sizes. Accurate results with any of the techniques are obtained only when measurements at a series of concentrations are extrapolated to infinite dilution where the system becomes ideal, i.e. is not affected by interactions between molecules.

Direct vapor pressure measurements with a differential manometer are generally limited to the larger depressions produced by low molecular weight solutes, while refined techniques such as isothermal distillation require the most exact control of conditions. Isopiestic methods allow the comparison of the vapor pressure of solutions of an unknown with those containing a known substance, and several modifications have been used more than other vapor pressure methods. Ebulliometric techniques which depend on the elevation of the boiling point of a solvent are often used for solutes of low molecular weight and find some use for large molecules. Since boiling points are highly sensitive to the atmospheric pressure, it is either necessary to control pressure very precisely, or more commonly to measure the boiling points of both the solvent and solution simultaneously. Often a differential thermometer is employed to determine only the difference of the two temperatures, and these devices have been made so sensitive that molecular weights as large as 30 000 have sometimes been studied. Techniques involving the lowering of the freezing point of a solvent (cryoscopic methods) are much used for rapid approximate determinations of molecular weights in the identification of organic compounds. For this purpose a substance such as camphor, which is a good solvent for many organic compounds and has a large molar depression constant, is often chosen to magnify the difference in freezing point of the solvent and the solution of the unknown. Since freezing-point depressions are not sensitive to atmospheric pressure, they are easier to measure accurately than the methods described above, and much use has been made of them for precise studies of solutes having low molecular weights. The possibility of association or ionization of the solute must be considered with any of these methods, since these effects will greatly influence the result.

Osmotic pressures are so much larger than any other colligative property that they are most widely used for molecular weight measurements,

particularly for long-chain polymers where the high sensitivity of the method is required. For accurate measurements, a membrane is required which permits the flow of solvent through its pores but completely holds back solute molecules. This condition is best satisfied where there are large differences in size of the solute and solvent molecules or of their affinity for the membrane. Membranes made from cellulose compounds are often successfully used for polymers which contain little material with molecular weights below about 10 000. Below this molecular weight the pore size of the satisfactory membranes is so small that solvent flow is very slow, and thus a very long time is required to reach constant osmotic pressure. In spite of this handicap, some of the most precise osmotic pressure measurements have been obtained with aqueous solutions of sucrose and similar small solutes by the use of membranes prepared by precipitating such materials as copper ferrocyanide in the pores of a solid support. The upper limit of molecular weights satisfactorily measured by osmometry is usually about 500 000, which is fixed by the lowest pressures that can be measured precisely and by the maximum concentrations of material which still give satisfactory extrapolations to infinite dilution. In comparing various colligative properties for the characterization of polymers, osmometry has the advantage that it is unaffected by the presence of impurities of very low molecular weight which will diffuse through membranes able to retain the polymer, whereas the other properties are greatly affected by the same impurities.

Modern instrumentation has provided commercial instruments utilizing several of these colligative properties for routine, accurate measurements in very short time and with small samples. This is true for boiling point, vapor pressure, and freezing point measurements of molecular weights up to several thousand, and for membrane osmotic pressure measurements of high molecular weight samples.

X-ray Diffraction. X-ray diffraction analysis is a powerful method for determining exact molecular weight and structural characteristics of compounds in their crystalline state. However, the method is complicated and slower than many techniques which provide molecular weights of accuracy sufficient for many purposes and so is usually employed only when the additional structural information is needed. The sample to be examined must have a high degree of crystalline order and is preferably a single crystal at least 0.1 mm in size; such samples are prepared fairly readily from many inorganic and non-polymeric organic compounds. Alternatively, crystalline powders of certain crystal types may provide suitable results. Diffraction patterns are then obtained by one of several methods, and the angular positions of the reflections are used to calculate the lattice spacings, and thus the size of the unit cell. This unit cell is the smallest volume unit which retains all geometrical features of the crystalline class, and it contains a small integral

number of molecules. A rough estimate of the molecular weight of the compound is needed from a determination by an independent method in order to obtain this integral number. Finally, the resultant molecular volume is multiplied by the exact bulk density of the crystal and by the avogadro number to yield the molecular weight (see X-RAY DIFFRACTION).

Light Scattering. Measurements of the intensity of light scattered by dissolved molecules allow the determination of molecular weights. Most commonly the method is used for polymers above 10 000 units, though under optimum conditions molecular weights as low as 1000 have been determined. Since the intensity scattered by a given weight of dissolved material is directly proportional to the mass of each molecule, a weight-average value of the molecular weight is obtained for a polydisperse system. An average dimension of the molecule can also be obtained by a study of the angular variation of scattered light intensity, provided some dimension of the molecules exceeds a few hundred angstroms. The interaction between dissolved molecules substantially affects the intensity of scattered light so that extrapolation to infinite dilution of data collected at several polymer concentrations is required. The method has been so well developed in the last decade that it is now probably the most used method for determining absolute molecular weights of polymers. It is more rapid than osmometry and provides information on sizes which is furnished by few other methods. The greatest problem encountered is in the removal of suspended large particles which otherwise would distort the angular scattering pattern of the solutions. This is rather easily accomplished by filtration in some cases, but it may be a formidable difficulty for particles which are highly solvated or are peptized by the molecules to be studied. Auxiliary information is required on the refractive index increment of the sample, i.e., the change in refractive index of the solvent produced by unit concentration of the sample. This information is supplied by a differential refractometer using the same wavelength of light as that employed in measurements of the intensity of scatter.

The Ultracentrifuge. The sedimentation of large molecules in a strong centrifugal field enables the determination of both average molecular weights and the distribution of molecular weights in certain systems. When a solution containing polymer or other large molecules is centrifuged at forces up to 250 000 times gravity, the molecules begin to settle, leaving pure solvent above a boundary which progressively moves toward the bottom of the cell. This boundary is a rather sharp gradient of concentrations for molecules of uniform size, such as globular proteins, but for polydisperse systems, the boundary is diffuse, the lowest molecular weights lagging behind the larger molecules. An optical system is provided for viewing this boundary, and a study as a function of the time of centrifuging yields the rate of sedimentation for the single

component or for each of many components of a polydisperse system. These sedimentation rates may then be related to the corresponding molecular weights of the species present after the diffusion coefficients for each species are determined by independent experiments. Both the sedimentation and the diffusion rates are affected by interactions between molecules, so that each must be studied as a function of concentration and extrapolated to infinite dilution as is done for the colligative properties. The result of this detailed work is the distribution of molecular weights in the sample which is available by few other methods. At present, this method is only partly satisfactory for molecular weight determinations with linear polymers because of the large concentration dependence of the diffusion coefficients. Difficulties have been found in reliably extrapolating diffusion coefficients beyond the lowest polymer concentrations which are experimentally attainable at present.

A modification of the sedimentation method which avoids the study of diffusion constants is the sedimentation equilibrium method in which molecules are allowed to sediment in a much weaker field. Under these conditions, the sedimenting force is balanced by the force of diffusion, so that after times from a day to two weeks molecules of each size reach different equilibrium positions, and the optical measurement of the concentration of polymer at each point gives the molecular weight distribution directly. However, again extrapolation to infinite dilution must be used to overcome interaction effects. The chief difficulty here is the long time of centrifuging required, and the necessary stability of the apparatus during the period.

A newer and somewhat faster technique, the Archibald method, permits the determination of weight-average molecular weights of polymers by analysis of the concentration gradient near boundaries soon after sedimentation begins.

Chemical Analysis. When reactive groups in a compound may be determined exactly and easily, this analysis may be used to determine the gram equivalent weight of the substance. This is the weight in grams which combines with or is equivalent to one gram-atomic weight of hydrogen. This equivalent weight may then be converted to the molecular weight by multiplying by the number of groups per molecule which reacted (provided they each are also equivalent to one hydrogen). If the number of reactive groups in the molecule is not known, then one of the physical methods for determining molecular weight must be used instead. The chemical method is convenient and often used for the identification of organic substances containing free carboxyl or amino groups which can readily be titrated, and for esters which can be saponified and determinations made of the amount of alkali consumed in this process. The equivalent weights of ionic substances containing, for example, halide or sulfate groups may also be determined by titration or by gravimetric analysis of insoluble compounds

formed with reagents which act in a stoichiometric fashion. In the titration of acids, the "neutral equivalent" is the weight of material which combines with one equivalent of alkali, and a similar definition applies to the "saponification equivalent" of esters. If only one carboxyl or ester group is present in the molecule, these values equal the molecular weight of the compound.

In a similar way, if the terminal groups on polymer chains can be determined by a chemical reaction without affecting other groups in the molecule, the equivalent weight or molecular weight of the polymer may be obtained in certain cases. For polydisperse systems, a number-average value of the molecular weight is obtained because the process essentially counts the total number of groups per unit weight of sample. Since the method depends on the effect of a single group in a long chain, its sensitivity decreases as the molecular weight rises, and so is seldom applicable above molecular weights of 20 000. Particularly at high molecular weights, the method is very sensitive to small amounts of impurities which can react with the testing reagent, so that careful purification of samples is desired.

It is also important to know that impurities or competing mechanisms of polymerization do not lead to branching or other processes which may provide greater or fewer reactive groups per molecule. The analysis for end groups must be carried out under mild conditions which do not degrade the polymer, since this would also lead to lower molecular weight values than expected. Labeling of end groups either with radioactive isotopes or with heavy isotopes which can be analyzed with the mass spectrometer provides a rapid and convenient analysis for end groups. This labeling can be accomplished with a labeled initiator if this remains at the chain ends, or after polymerization is complete, by exchange of weakly bonded groups with similar groups in a labeled compound. Molecular weight determinations by end group analysis are often used for condensation polymers of lower molecular weights and are especially valuable in studying degradation processes in polymers.

GEORGE L. BEYER

References

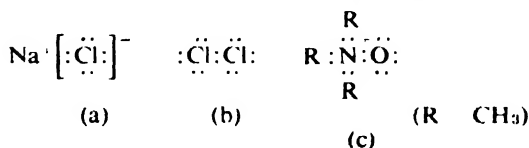
- Daniels, F., Williams, J. W., Bender, P., Alberty, R. A., and Cornwell, C. D., "Experimental Physical Chemistry," Sixth edition, New York, McGraw-Hill Book Co., 1962.
- Bonnar, R. U., Dimbat, M., Stross, F. H., "Number-Average Molecular Weights," New York, Interscience Publishers, 1958.
- Allen, P. W., "Techniques of Polymer Characterization," London, Butterworths, 1959.
- Stacey, K. A., "Light-Scattering in Physical Chemistry," New York, Academic Press, 1956.
- Wells, A. F., "Structural Inorganic Chemistry," Oxford, Oxford University Press, 1945.

Cross-references: ATOMIC PHYSICS; BOND, CHEMICAL; CENTRIFUGE; LIGHT SCATTERING; MOLECULES AND MOLECULAR STRUCTURE; OSMOSIS; POLYMER PHYSICS; VAPOR PRESSURE AND EVAPORATION; X-RAY DIFFRACTION.

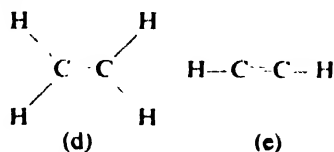
MOLECULES AND MOLECULAR STRUCTURE

A molecule is a local assembly of atomic nuclei and electrons in a state of dynamic stability. The cohesive forces are electrostatic, but, in addition, relatively small electromagnetic interactions may occur between the spin and orbital motions of the electrons, especially in the neighborhood of heavy nuclei. The internuclear separations are of the order of 1 to 2×10^{-10} metres, and the energies required to dissociate a stable molecule into smaller fragments fall into the 1 to 5 eV range. The simplest diatomic species is the hydrogen molecule-ion H_2^+ with two nuclei and one electron. At the other extreme, the protein ribonuclease contains 1876 nuclei and 7396 electrons per molecule.

Historically, molecules were regarded as being formed by the association of individual atoms. This led to the concept of *valency*, i.e., the number of individual chemical bonds or linkages with which a particular atom can attach itself to other atoms. When the electronic theory of the atom was developed, these bonds were interpreted in terms of the behavior of the valence, or outer shell, electrons of the combining atoms. Each atom with a partly filled valence shell attempts to acquire a completed octet of outer electrons, either by electron transfer, as in (a), to give an electrovalent bond, resulting from Coulombic attraction between the oppositely charged ions

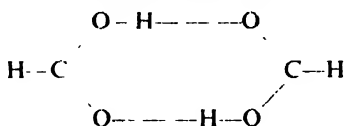


(W. Kossel, 1916); or by electron sharing, as in (b) and (c), to give a covalent bond (G. N. Lewis, 1916). In (b), each chlorine atom donates one electron to form a *homopolar* bond, which is written $\text{Cl}-\text{Cl}$ where the bar denotes on this theory one single bond, or shared electron pair. In (c), the nitrogen-oxygen bond is formed by two electrons donated by only the nitrogen atom, giving a *semipolar*, or *coordinate-covalent* bond, which is written $\text{R}_3\text{N}^+-\text{O}^-$, and which is electrically polarized. Double or triple bonds result from the sharing of four or six electrons between adjacent atoms, as in ethylene (d) and acetylene (e) respectively.



However, difficulties arise in describing the structures of many molecules in this fashion. For example, in benzene (C_6H_6), a typical aromatic compound, the carbon nuclei form a plane regular hexagon, but the electrons can only be conventionally written as forming alternate single and double bonds between them. Furthermore, an electron cannot be identified as coming specifically from any one of these bonds upon ionization. Such difficulties disappear in the quantum-mechanical theory of a polyatomic molecule, whose electronic wave function can be constructed from nonlocalized electron orbitals extending over all of the nuclei. The concept of valency is not basic to this theory, but is simply a convenient approximation by which the electron density distribution is partitioned in different regions in the molecule.

Molecular compounds consist of two or more stable species held together by weak forces. In *clathrates*, a gaseous substance such as SO_2 , HCl , CO_2 or a rare gas is held in the crystal lattice of a solid, such as β -quinol, by van der Waals-London dispersion forces. The *gas hydrates*, e.g., $Cl_2 \cdot 6H_2O$, contain halogen molecules similarly trapped in ice-like structures. The hydrogen bond, with energy ~ 0.25 eV, is responsible not only for the high degree of molecular association in liquids such as water ($O-H \cdots O-H \cdots$) but also for such molecules as the formic acid dimer



which contains two hydrogen bonds indicated by dashed lines. *Molecular complexes* vary greatly in their stability; in donor-acceptor complexes, electronic charge is transferred from the donor (e.g., NH_3) to the acceptor (e.g., BF_3), as in a semipolar bond. The $BF_3 \cdot NH_3$ complex has a binding energy with respect to dissociation into NH_3 and BF_3 of 1.8 eV. The bond here is relatively strong; the electron transfer can occur between the components in their electronic ground states. On the other hand, in weaker complexes such as $C_6H_6 \cdots I_2$, with binding energy of about 0.06 eV, there is only a fractional transfer of charge from benzene to iodine. The actual ionic charge-transfer state lies at much higher energy than the ground state of the complex.

The discovery of $XePtF_6$ by Bartlett (1962) has been followed by the synthesis of many other rare gas compounds whose existence was not predicted by classical valency theories. Compounds such as XeF_2 , XeF_4 , XeF_6 and $XeOF_4$ are quite stable, the average $Xe-F$ bond energy in the square planar molecule XeF_4 being 1.4 eV.

A molecule X is characterized by:

(1) A *stoichiometric formula* $A_aB_bC_c \cdots$ where a, b, c, \dots are the numbers of atoms of elements A, B, C, \dots that it contains. The ratio $a : b : c : \dots$ is found by chemical analysis for these elements.

The absolute values of a, b, c, \dots are then fixed by determination of the *molecular weight* of X . For a volatile substance, the gas density of X and of a gas of known molecular weight are compared at the same temperature and pressure. The molecular weights are in the ratio of the gas densities, since *Avogadro's principle* states that equal volumes of gases at the same temperature and pressure contain the same numbers of molecules. For a nonvolatile substance, a known weight can be dissolved in a solvent, and the resultant lowering of vapor pressure, elevation of the boiling point, or depression of the freezing point of the solvent can be measured. Each of these properties depends upon the number of molecules of solute present, so the number of molecules per unit weight of X is found and, hence, the molecular weight. For substances of high molecular weight such as proteins (molecular weight $\sim 34\,000$ – $200\,000$) or polymers, the molecular weight is found from osmotic pressure measurements or the rate of sedimentation in a centrifuge. The molecular weight of a molecule in crystalline form is determined when the density of the crystal and the dimensions of the unit cell from x-ray analysis are both known. Finally, for stable volatile compounds, it is often possible to form the ion X^+ and pass this through a mass spectrograph to determine the molecular weight.

(2) The spatial distribution of the nuclei in their mean equilibrium or "rest" positions. At an elementary level, this is described in geometrical language. For example, in carbon tetrachloride, CCl_4 , the four chlorine nuclei are disposed at the corners of a regular tetrahedron, and the carbon nucleus is at the center. In the $[CoCl_4]^{2-}$ ion, the arrangement of the chlorine nuclei about the central metal nucleus is also tetrahedral, whereas in $[PdCl_4]^{2-}$ it is planar.

At a more sophisticated level, each molecule is classified under a *symmetry point group*. Most nonlinear molecules possess only 1, 2, 3, 4 or 6-fold rotation axes, and belong to one of the 32 crystallographic point groups. For example, the pyramidal ammonia molecule NH_3 has a three-fold rotation axis C_3 through the nitrogen nucleus and three reflection planes σ_v intersecting at this axis, and belongs to the $C_{3v}(3m)$ point group. Tetrahedral molecules CX_4 belong to the $T_d(\bar{4}3m)$ point group. Linear diatomic and polyatomic molecules belong to either of the continuous point groups $D_{\infty h}$ or $C_{\infty v}$ according to whether a center of symmetry is present or not.

The symmetry classification does not define the geometry of a molecule completely. The values of certain *bond lengths or angles* must also be specified. In carbon tetrachloride, it is sufficient to give the $C-Cl$ distance (1.77×10^{-10} meters) since classification under the T_d point group implies that all four of these bonds have equal length and the angle between them is $109^\circ 28'$. In ammonia, both the $N-H$ distance (1.015×10^{-10} meters) and the angle HNH (107°) must be specified. In general, the lower the molecular symmetry, the greater is the number of such

independent parameters required to characterize the geometry. Information about the symmetry and internal dimensions of a molecule is obtained experimentally by SPECTROSCOPY, ELECTRON DIFFRACTION, NEUTRON DIFFRACTION and X-RAY DIFFRACTION.

(3) The *dynamical state* is defined by the values of certain observables associated with orbital and spin motions of the electrons and with vibration and rotation of the nuclei, and also by symmetry properties of the corresponding stationary-state wave functions. Except for cases when heavy nuclei are present, the total electron spin angular momentum of a molecule is separately conserved with magnitude Sh , and molecular states are classified as singlet, doublet, triplet, ... according to the value of the multiplicity $(2S+1)$. This is shown by a prefix superscript to the term symbol, as in atoms.

The Born-Oppenheimer approximation permits the molecular Hamiltonian H to be separated into a component H_e that depends only on the coordinates of the electrons relative to the nuclei plus a component depending upon the nuclear coordinates, which in turn can be written as a sum $H_v + H_r$ of terms for vibrational and rotational motion of the nuclei (we may ignore translation here). The eigenfunctions Ψ of H may correspondingly be factorized as the product $\Psi_e \Psi_v \Psi_r$ of eigenfunctions of these three operators, and the eigenvalues E decomposed as the sum $E_e + E_v + E_r$. In general, we find $E_e \gg E_v > E_r$.

Electronic states of molecules are classified according to the symmetry properties of Ψ_e (which forms a basis for an irreducible representation of the molecular point group). Thus ${}^3B_{1u}$ is a term symbol for benzene (D_{6h} point group) that denotes a triplet electronic state whose wave function transforms like the B_{1u} representation of the group. In the case of diatomic and linear polyatomic molecules, the term symbol shows the magnitude of the conserved component of orbital electronic angular momentum Λh about the axis, states being classified as Σ , Π , Δ , ... according to $\Lambda = 0, 1, 2, \dots$. The superscript $+$ or $-$ shows the behavior of Ψ_e for a linear molecule upon reflection in a plane containing the molecular axis; for centrosymmetric linear molecules ($D_{\infty h}$ point group) the subscript g or u shows the parity $+1$ or -1 respectively for Ψ_e with respect to inversion at the center.

The vibrational wavefunction Ψ_v can be approximated by a product of $3N-6$ harmonic oscillator wave functions ψ_i , each a function of a normal displacement coordinate Q_i ,

$$\Psi_v = \prod_{i=1}^{3N-6} \psi_i(Q_i)$$

The product is $(3N-5)$ for a linear molecule; N is the number of nuclei. Each oscillatory mode can be excited with quanta $v_i = 0, 1, 2, \dots$. When $v_i = 0$, ψ_i transforms like the totally symmetrical representation of the molecular point group;

when $v_i = 1$, ψ_i transforms like Q_i . The symmetry of Ψ_v under the molecular point group is found from the direct product for all the ψ_i . The vibrationless ground state with $v_1 = v_2 = \dots = 0$ is always totally symmetrical.

Each rotational state is characterized by a value for the quantum number J , where $J(J+1)\hbar^2$ is the squared angular momentum for rotation of the nuclei (apart from spin). If I_a , I_b and I_c denote the moments about the principal axes of inertia of the molecule, then a spherical top has $I_a = I_b = I_c$; a molecule with two principal moments equal is either a prolate ($I_c = I_b > I_a$) or an oblate ($I_c > I_b = I_a$) symmetric top; if $I_c \neq I_b \neq I_a$, the top is asymmetric. Symmetric top molecules have C_n symmetry axes with $n \geq 3$ and belong to point groups with degenerate representations. The component $K\hbar$ of rotational angular momentum about the top axis is conserved and the rotational levels are also characterized by the value of the quantum number $K = 0, 1, 2, \dots, J$. A symmetry classification is made for Ψ_r under the rotational subgroup of the molecular point group. Finally, each eigenstate is described as $+$ or $-$ according to the parity of Ψ under inversion in a space-fixed coordinate system.

(4) In order to distinguish between different electronic states Ψ_e of the same symmetry and spin multiplicity, a further classification is obtained by expanding Ψ_e as a product of n single-electron wave functions ϕ_i , each a function of the coordinates of one of the n electrons in the molecule.

$$\Psi_e = (n!)^{-1/2} \det[\phi_1(1)\phi_2(2)\phi_3(3) \dots \phi_n(n)]$$

where $(n!)^{-1/2}$ is a normalization factor. Each of the molecular orbitals (MO's) ϕ_i is constructed to transform like an irreducible representation of the molecular point group and is usually formed by linear combination of atomic orbitals (LCAO) χ_i centered upon the individual nuclei

$$\phi_i = \sum_p C_{ip} \chi_p$$

The MO's are written in order of decreasing energy necessary to ionize the electrons which occupy them, and electrons are assigned to the MO's in accordance with the Pauli principle. For example, the electronic ground state of ammonia (C_{3v} point group) is written

$$(1a_1)^2(2a_1)^2(1e)^4(3a_1)^2 \quad {}^1A_1$$

where the superscripts show the distribution of the ten electrons among three MO's of a_1 symmetry and one of e symmetry, the electrons in the $(3a_1)$ orbital being most readily ionized. The symmetry of the resultant molecular wavefunction Ψ_e is found by taking direct products for each orbital occupied by an electron. Here Ψ_e belongs to the totally symmetrical representation (and is also singlet). Excited electronic states are obtained by promoting electrons into orbitals with higher energies, but the molecular symmetry in such

states often differs from that in the ground state, as a result of changes in geometry.

G. W. KING

References

- Coulson, C. A., "Valence," Second edition, London, Oxford University Press, 1961.
 Eyring H., Walter, J., and Kimball, G. E., "Quantum Chemistry," New York, John Wiley & Sons, 1944.
 Ketelaar, J. A. A., "Chemical Constitution," Second edition, Amsterdam, Elsevier Publishing Co., 1958.
 King, G. W., "Spectroscopy and Molecular Structure," New York, Holt, Rinehart and Winston, Inc., 1964.
 Pauling, L. C., "The Nature of the Chemical Bond," Third edition, Ithaca, N.Y., Cornell University Press, 1960.

Cross-references: BOND, CHEMICAL; ELECTRON DIFFRACTION; MOLECULAR WEIGHT; NEUTRON DIFFRACTION; QUANTUM THEORY; SPECTROSCOPY; X-RAY DIFFRACTION.

MOMENTUM. See IMPULSE AND MOMENTUM.

MÖSSBAUER EFFECT

The Mössbauer effect is the phenomenon of recoilless resonance fluorescence of gamma rays from nuclei bound in solids. It was first discovered in 1958 and brought its discoverer, Rudolf L. Mössbauer, the Nobel prize for physics in 1961. The extreme sharpness of the recoilless gamma transitions and the relative ease and accuracy in observing small energy differences make the Mössbauer effect an important tool in nuclear physics, solid-state physics and chemistry.

Resonance fluorescence involves the excitation of a quantized system (the absorber) from its ground state (0) to an excited state (1) by absorption of a photon emitted from an identical system (the source) decaying from state (1) to (0). The parameters characterizing the nuclear resonance process for some typical isotopes are illustrated in Fig. 1.

To conserve energy and momentum in the emission and absorption processes, each system, the source and absorber, must acquire a recoil energy R equal to $E^2/2Mc^2$, where E is the photon energy. M is the mass of the recoiling system and c is the speed of light. The energy available for the excitation of the absorber is thus reduced by $2R$, and resonance fluorescence can be achieved only if the missing energy $2R$ is not larger than the widths of the levels involved. Before 1958, it was thought that for all gamma transitions the width required to get overlap between the emission and the absorption line was much larger than the natural width Γ , where Γ is related to the half-life $T_{1/2}$ of the excited nuclear level by the expression $\Gamma T_{1/2} = 4.55 \times 10^{-16} \text{ eV sec}$. In fact, techniques had been developed to compensate for the recoil energy loss by applying large Doppler shifts with an ultracentrifuge or through

thermal motion. These methods necessarily broaden the intrinsically narrow lines thereby reducing the absorption cross section.

Modifying a theory of W. E. Lamb, Mössbauer demonstrated that in some cases these difficulties may be removed by embedding the source and absorber nuclei in a crystal. Being part of a quantized vibrational system, these nuclei interact with the lattice by exchange of vibrational quanta or phonons only. If the characteristic phonon energy is large compared to the recoil energy R for a free nucleus, the probability for the emission of a gamma ray without a change in the vibrational state of the lattice is large. For such a zero phonon transition, the lattice as a whole absorbs the recoil momentum and the recoil energy loss is negligibly small. At the same time, the emission and absorption lines achieve the natural width Γ .

For an atom bound by harmonic forces, the fraction f of events without recoil energy loss is given by $f = \exp(-4\pi^2 \langle x^2 \rangle / \lambda^2)$. Here $\langle x^2 \rangle$ is the mean square displacement of the radiating atom taken along the direction of the photon with wavelength λ . In an environment of lower than cubic symmetry, $\langle x^2 \rangle$, and therefore f , may be anisotropic. A large recoilless fraction may be obtained when $\langle x^2 \rangle$ is small and λ large. The former condition implies small vibrational amplitude and thus low temperature, high vibrational frequency and large mass M , while the latter implies low photon energy, E . Both conditions imply small recoil energy R .

Recoilless transitions can also occur in amorphous substances like glasses and high-viscosity liquids. For the latter, the diffusive motion superimposed on the thermal vibration results in a broadening of the Mössbauer line.

For all Mössbauer isotopes, the nuclear half-life $T_{1/2}$, typically 10^{-8} second, is very long compared to the period of the lattice vibrations, typically 10^{-11} second. A conceivable first-order Doppler shift of the Mössbauer line due to the thermal motion will therefore average out to zero. The second-order Doppler effect, however, leads to an observable shift, sometimes called the temperature shift. The photons emitted by a source nucleus moving with a mean square velocity $\langle v^2 \rangle$ are lower in energy by a fraction $\langle v^2 \rangle / 2c^2$ as compared to the photons emitted at rest. Similarly the transition energy of a moving absorber nucleus appears lower to the incident photon by a fraction $\langle v^2 \rangle / 2c^2$. In principle, the two shifts may be different whenever the source and absorber are of different composition and/or temperature.

Mössbauer performed his original experiment with Ir^{191} at 88 K, obtaining a recoilless fraction of 1 per cent. Since the natural line width in Ir^{191} , as in most other Mössbauer nuclides, is extremely narrow, Mössbauer was able to alter the degree of overlap between the emission and absorption lines by simply moving the source relative to the absorber at speeds v of the order of 1 mm/sec. Thus the gamma rays were slightly shifted in energy via the first-order Doppler effect by an

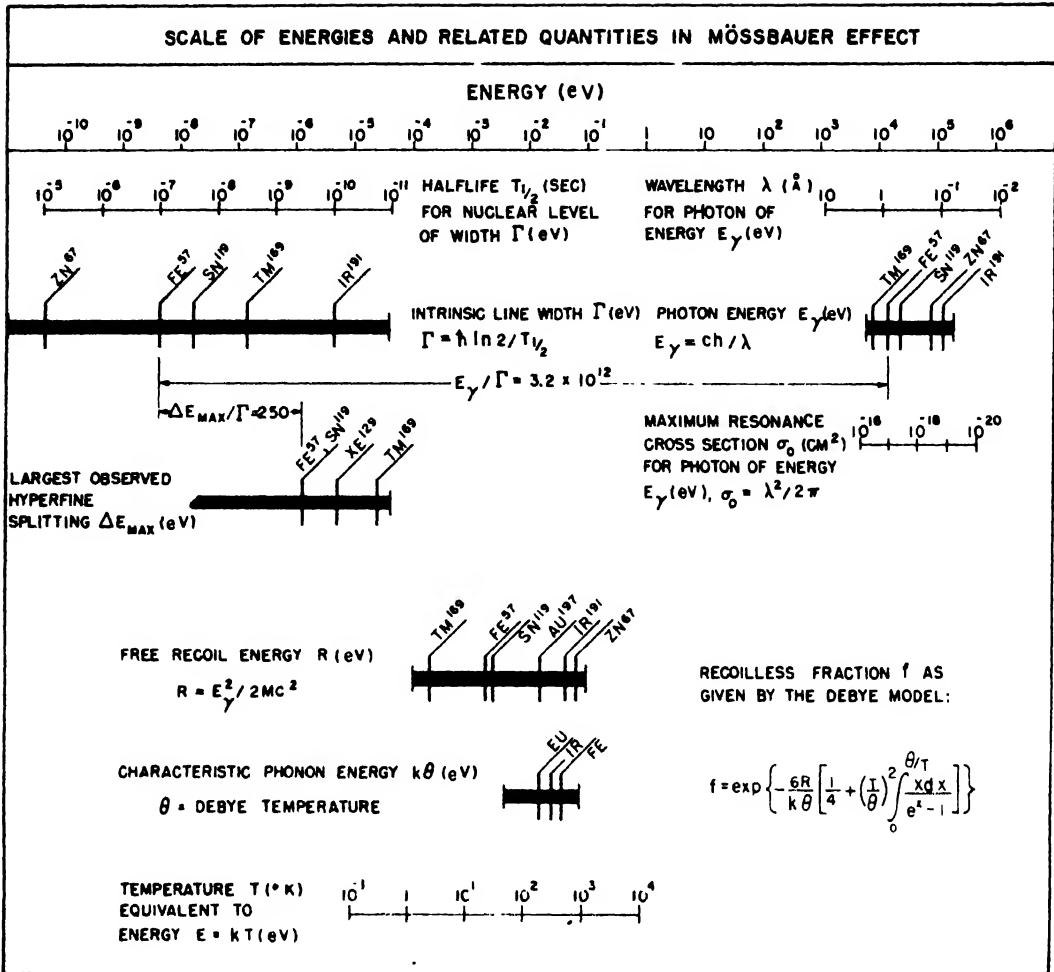


FIG. 1

amount $\Delta E = Ev/c$. By plotting the transmission through the absorber as a function of the relative source-absorber velocity, one thus obtains the characteristic Mössbauer velocity spectrum which exhibits the shape of the resonance curve. From such a plot, one can determine the recoilless fraction, the lifetime of the excited state and any possible energy differences between the emission and the absorption line.

With extreme care, it is possible to determine energy differences of the order of $1/1000$ of the line width Γ . The latter typically varies with isotope from 10^{-10} to 10^{-15} times the actual gamma ray energy E . The Mössbauer effect therefore enables one to detect extremely small changes in this energy. One of the earliest applications of this great precision was the laboratory verification of the gravitational red shift by Pound and Rebka. According to Einstein's theory, photons have an apparent mass $m = E/c^2$. Thus if they fall toward the earth through a distance H , their energy increases by $\Delta E = mgH$, so that

$\Delta E/E = gH/c^2 \approx 10^{-16}$ per meter. Using Fe^{57} , which has a large recoilless fraction, and for which $\Gamma/E \approx 3 \times 10^{-13}$, the desired effect was observed when the photons were sent down the 22-meter tower at Harvard University.

It is well known from optical and high-frequency spectroscopy that a nucleus interacting with its environment through its charge distribution and magnetic moment can give rise to hyperfine shifts and splittings of the order of 10^{-6}eV to 10^{-9}eV . In Mössbauer experiments, such energy differences can readily be measured since the line width of the recoilless transitions is of the same order of magnitude. Perhaps, therefore, the most useful feature of the Mössbauer effect is that it may be used to obtain nuclear properties if the fields acting on the nucleus are known, and conversely, it is a powerful tool for probing solids once the various interactions are calibrated, i.e., the nuclear properties have been determined. Some representative results obtained with Fe^{57} are illustrated in Fig. 2.

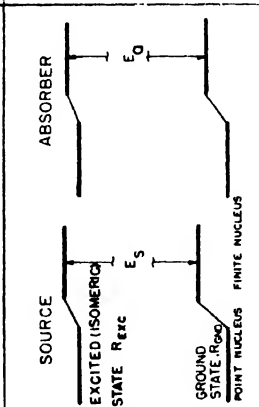
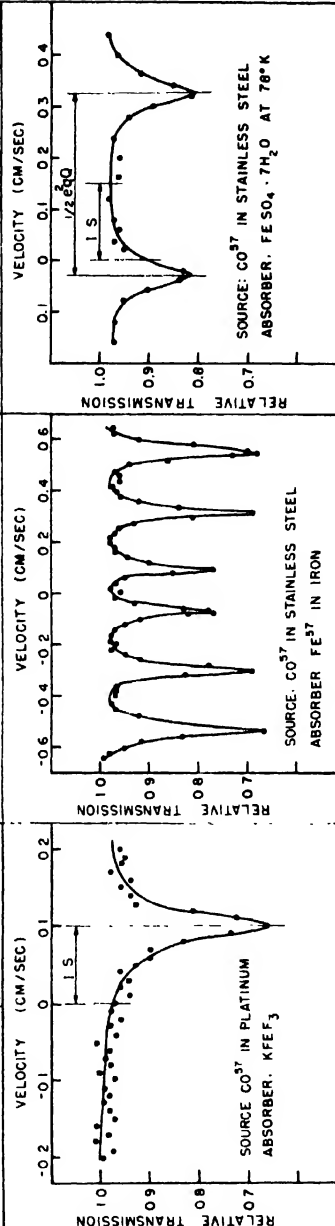
NUCLEAR HYPERFINE INTERACTION FOR Fe^{57}					
MULTIPOLE ORDER	ELECTRIC MONOPOLE · ISOMER SHIFT	MAGNETIC DIPOLE : ZEEMAN SPLITTING	ELECTRIC QUADRUPOLE		
NUCLEAR PROPERTY	CHANGE IN CHARGE RADIUS $\frac{\delta R}{R}$	MAGNETIC MOMENT μ	ELECTRIC QUADRUPOLE MOMENT Q		
ATOMIC PROPERTY	s - ELECTRON DENSITY $ \psi(o) ^2$	INTERNAL MAGNETIC FIELD $H(o)$	ELECTRIC FIELD GRADIENT q		
INTERACTION ENERGY	$IS_0 = E_0 - E_s = \frac{4\pi}{5} Ze^2 R^2 \left(\frac{\delta R}{R} \right) \psi(o) ^2 - \psi(o) ^2$ $E_M = \frac{\mu H(o) I_z}{I}$ $E_Q = eqQ \frac{3I^2 - I(I+1)}{4I(2I-1)}$				
ENERGY LEVEL DIAGRAM WITH GAMMA TRANSITIONS ALLOWED BY SELECTION RULES	 <p>EXAMPLE: Fe^{57} IN PT VS Fe^{57} IN $KFeF_3$ $\frac{\delta R}{R} = \frac{R_{exc} - R_{gnd}}{R} = 0.0018, \psi(o) ^2 = 1.7 \times 10^4$ $\frac{\delta R}{R} = \frac{R_{gnd}}{R} = 0.0018$</p>				
TYPICAL MÖSSBAUER SPECTRA	 <p>SOURCE: Co^{57} IN PLATINUM ABSORBER: $KFeF_3$</p> <p>SOURCE: Co^{57} IN STAINLESS STEEL ABSORBER Fe^{57} IN IRON</p> <p>SOURCE: Co^{57} IN STAINLESS STEEL ABSORBER: $FeSO_4 \cdot 7H_2O$ AT $78^\circ K$</p>				

FIG. 2.

The most basic of these interactions is the effect of the finite nuclear size which, in general, is different for the ground state and the excited state. The electrostatic interaction of the nuclear charge with the s -electrons overlapping it raises the nuclear energy levels by an amount depending on the charge radii and s -electron density at the nucleus. Therefore under proper conditions, there appears a shift in the Mössbauer resonance, the isomer shift, which is proportional to $(\delta R/R) \delta|\psi(0)|^2$, where $\delta R/R$ is the fractional change in the nuclear radius during the decay and $\delta|\psi(0)|^2$ is the difference in s -electron density between source and absorber. To determine the quantity $\delta R/R$ one compares the isomer shifts of two chemically simple absorbers, for which the s -electron density can be calculated. In the case of Fe^{57} an isomer shift exists between compounds containing ferric ions, $\text{Fe}^{3+}(3d^5)$, and ferrous ions, $\text{Fe}^{2+}(3d^6)$. Although the number of s -electrons is the same for both ions, a detailed calculation shows that the shielding through the additional $3d$ electron changes the $3s$ density at the nucleus. For I^{129} , the isomer shifts observed among different alkali iodides can be related quantitatively to the known transfer of $5p$ electrons to the ligands which affects the $5s$ density at the nucleus. Once calibrated, the isomer shift is a tool for measuring s -electron densities and is therefore of use in studying chemical bonding, energy bands in solids, and also in identifying charge states of a given atom.

One of the early successes of the Mössbauer effect was the observation of the completely resolved nuclear Zeeman splitting arising from the magnetic hyperfine interaction of Fe^{57} in ferromagnetic iron. For this isotope, as well as for most other Mössbauer isotopes, the magnetic moment of the nuclear ground state is known from magnetic resonance experiments, and the calibration is therefore straightforward. Careful analysis of the velocity spectrum for magnetic samples is sufficient in general to reveal both the desired magnetic moment and internal magnetic field. The latter yields important information about the unpaired spin density at the nucleus, which in turn is related to the exchange interaction in crystals, metals and alloys. For single crystals or magnetized samples, the intensities of the individual lines of the Mössbauer spectrum depend on the angle between the direction of the internal field and the emitted photon. From a measurement of the intensity distribution, one therefore obtains the direction of the internal magnetic field. The temperature dependence of the splitting can yield Néel and Curie temperatures and also relaxation times.

Whenever one of the nuclear levels possesses a quadrupole moment and an electric field gradient exists at the position of the nucleus, quadrupole splitting of the Mössbauer spectrum may be observed. If the quadrupole moment is known either for the ground state or for the excited state, then a Mössbauer measurement will readily yield the parameters of the field gradient tensor. Usually, however, the quadrupole moment is not known, and the field gradient tensor must be

determined from other work or else calculated from first principles. This tensor exists whenever the symmetry of the surrounding charge distribution is lower than cubic, and it is generally specified by two independent parameters. This tensor is easiest to calculate for cases of axial symmetry, in which it is characterized by one parameter, the field gradient, q . For simple ionic systems, it is possible to estimate q with some degree of certainty, and thereby determine the quadrupole moment. Once this is done, the Mössbauer effect may be used to measure field gradient tensors in more complicated systems. Such measurements yield information about crystalline symmetries, crystalline field splittings, shielding due to closed shell electrons, relaxation phenomena and chemical bonding. In addition, with single crystals, a study of the relative intensity of the various lines of the resonance spectrum as a function of angle can yield information about the orientation of the crystalline field axes and, thus, the orientation of complexes in solids. In cases where both magnetic and quadrupole splitting are present, the analysis becomes more complicated since it depends markedly upon the relative angle between the magnetic field and the axes of the electric field gradient tensor. Such cases, however, have been successfully handled for a number of antiferromagnetic compounds.

This article has only covered the basic features of the Mössbauer effect and the phenomena which affect the Mössbauer velocity spectrum in a general way. The actual application of the effect is extremely far reaching, embracing not only almost all areas of physics but also the fields of chemistry, biology, metallurgy and engineering. The reader is advised to consult the references for more information.

R. INGALLS
P. DEBRUNNER

References

- Mössbauer, R. L., *Science*, **137**, 731 (1962).
- Frauenfelder, H., "The Mössbauer Effect," New York, W. A. Benjamin, 1962.
- Boyle, A. J. F., and Hall, H. E., *Rept. Progr. Phys.*, **25**, 441 (1962).
- Wertheim, G. K., "Mössbauer Effect: Principles and Applications," New York, Academic Press, 1964.

Cross-references: CONSERVATION LAWS AND SYMMETRY, DOPPLER EFFECT, ISOTOPES, LUMINESCENCE, PHONONS, RADIOACTIVITY, ZEEMAN AND STARK EFFECTS.

MOTORS, ELECTRIC

Historical. Power conversion was discovered by M. Faraday in 1831; the commutator, by J. Henry, Pixii, and C. Wheatstone (1841); the electromagnetic field, by J. Brett (1840), Wheatstone and Cooke (1845), and W. von Siemens (1867); drum armatures, by Siemens, Pacinotti, and von Alteneck; ring armatures, by Gramme (1870); and disc armatures, by Desrozières (1885)

and Fritsche (1890). Ring and disk types are now seldom used. Revolving magnetic fields (1885) and ac theory were discovered by G. Ferraris; polyphase motors and systems, by N. Tesla (1888); the squirrel-cage rotor, by C. S. Bradley (1889); and ac commutator motors, by R. Eickemeyer, E. Thomson, L. Atkinson, and others.

Principles. These are the laws of Ohm, Kirchhoff, Lenz, and Maxwell; more specifically:

(1) Moving a conductor of length l across a magnetic field of flux of density B with a velocity v generates in it an electromotive force (emf) $e = vBl \times 10^{-8}$ volts. In motors, e opposes the current i and decreases with increase in load.

(2) The force on such a conductor equals $F = 0.1 Bil$ dynes. Fig. 1 shows the directions of current and force.

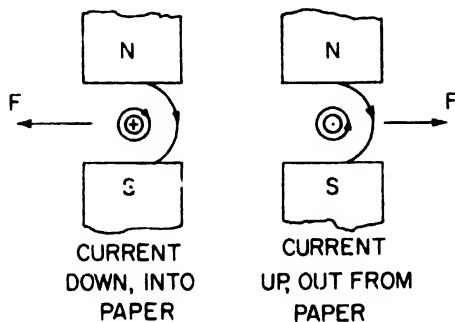


FIG. 1. Direction of force due to current in a magnetic field

(3) Magnetic structures tend to move to the position of minimum reluctance with a force $F = \frac{dW}{dx}$ where W is stored magnetic energy and x is distance.

(4) The force between two coupled circuits equals

$$F = \frac{i_1^2}{2} \frac{dL_1}{dx} + i_1 i_2 \frac{dM}{dx} + \frac{i_2^2}{2} \frac{dL_2}{dx}$$

Frequently L_1 and L_2 are constant and then $F = i_1 i_2 \frac{dM}{dx}$. Here i_1 and i_2 are currents; L_1 and L_2 are total self-inductances; M is mutual inductance.

Motor Types. Of many hundreds, the most used are:

(1) *Direct-Current.* (a) *Series:* The field coils are of heavy wire in series with the armature. The torque and current are high at low speeds and low at high speeds. These motors will run away at light loads.

(b) *Shunt:* The field coils of fine wire are in parallel with the armature. The speed drops slightly and current and torque increase with load. In very small-size motors (and sometimes in fractional and integral horsepower sizes up to $7\frac{1}{2}$ hp), permanent magnet fields are used.

(c) *Compound:* Both shunt and series coils are

used on the same motor. Behavior is intermediate between (a) and (b).

The current in all dc motors satisfies the relation $I_a = \frac{V_a - E_a}{R_a}$, where $E_a = \frac{p\phi_a Z_a S}{m_a \times 60 \times 10^8}$; V_a = terminal emf; E_a = the counter electromotive force (cemf); R_a = resistance; m_a = number of paths; and Z_a = number of conductors, all for the armature; p = number of poles, S = rpm, ϕ_a = useful flux per pole. ϕ_a may be nearly constant or it may be a function of I_a .

Speed control is achieved by adjusting field, current and/or armature terminal emf, tapped series coils, or (more rarely) field reluctance or double commutator. To avoid damage in starting, either starting resistances or voltage controls are needed, except for very small-size motors.

(2) *Alternating Current.* (a) *Polyphase Induction:* These are usually 3-phase, with phase windings spaced equally in slots around the periphery to produce alternate N and S poles. When fed with polyphase current, there is set up a revolving field which turns with the speed $S = \frac{120f}{p}$ rpm and induces the cemf $E = 2.22f\phi_m Zk_w \times 10^{-8}$ volts per phase, where ϕ_m is the flux per pole and k_w is the winding factor.

With rotor open or running at synchronous speed S , each phase behaves as an inductance and draws an exciting current lagging nearly $\frac{1}{2}$ cycle behind the emf. Shaft load reduces the speed from S to $S(1 - s)$, where s is slip. When referred to primary, s generates in each rotor phase the emf sE and current

$$I_2 = \frac{sE}{\sqrt{r_2^2 + (sx_2)^2}} = \frac{E}{\sqrt{\left(\frac{r_2}{s}\right)^2 + x_2^2}}$$

where r_2 is the resistance and x_2 is the leakage reactance of rotor in ohms per phase.

The current I_2 produces the needed torque. It introduces a new magnetomotive force (mmf) which turns at synchronous speed S in the same direction as that of the stator and lags behind it in space by the electrical angle $90^\circ + \tan^{-1} \frac{x_2}{r_2}$.

To balance this mmf, the stator draws an additional component of current sufficient to carry the load. Performance can be found (as for a trans-

former) if $\frac{r_2}{s}$ is taken as variable. Three-phase motors require less material than one or two-phase motors.

(b) *Single-phase Induction:* With the rotor at rest, there is no revolving field; the motor will start in either direction only if given a push. Rotation sets up an elliptical revolving field which turns synchronously at nonuniform speed in the same direction as the rotor and pulsates between the limits ϕ_m and $\phi_m(1 - s)$. The stator current is the resultant of (stator + cross-axis) magnetizing + load + loss currents. Analysis of this type of motor is less simple than that for a polyphase motor.

Starting is with a line switch, reduced voltage (auto-transformer or "compensator," wye-delta, series resistors or chokes), wound rotors and resistors, or more elaborate schemes. For single-phase motors, auxiliary start windings (split-phase or capacitor), or repulsion-start, are used and are cut out at about $\frac{1}{2}$ speed. For sizes 0 to $\frac{1}{2}$ hp, and sometimes for sizes up to 1 hp, shaded poles are used. *Speed Control* is by reduced voltage, wound rotor, or any one of a great number of possible techniques. Induction motors are "constant speed" machines and control apparatus is more expensive than that for direct current.

(c) Synchronous: Commonly these have a stationary phase-wound armature and revolving dc field structure. When up to speed and synchronized, the revolving mmf set up by the armature stands still relative to the dc field. When the angle $(I_a, E_a) = 0$, the armature mmf poles stand midway between the field poles. When I_a lags behind E_a , the armature mmf assists the dc field coils; it opposes them when I_a leads E_a . Armature current I_a takes a value and time phase position such that the cemf E_a and current I_a are correct to carry the existing load. Important relations are $I_a = \frac{V - E_a}{Z_a}$, power converted $= mI_a E_a \cos (I_a, E_a)$, input $= mI_a V \cos (I_a, V)$, and armature copper loss $= mI_a^2 r_a$, where E_a is a function of I_a , I_f , and angle (I_a, E_a) . In general, V and E_a are prevailing in opposition and I_a increases with load.

Synchronous motors are used where the rpm must be fixed: for power factor correction, regulating transmission line voltages, or speeds too low for good induction motor performance. They are not self-starting unless special means are provided, usually squirrel-cage or phase windings in the pole faces or part windings in combination with clutches (shaft or band brake on revolving stator). Precautions against high ac voltages in the dc field coils are needed when starting. The motors will carry some load with the dc field winding open. Small sizes (reluctance types) operate without field windings. Single-phase motors are less satisfactory because of lower efficiency, tendency to severe hunting, and problems in starting.

(d) ac Commutator: These are mostly single-phase series, repulsion, or combinations of these types. However, shunt and polyphase series, shunt compound, repulsion, and adjustable speed induction motors have been built. Single-phase series motors are used in large quantities for portable tools, vacuum sweepers, household appliances, etc., and electric railways; though for this last, repulsion and combination types are also used. Rectifiers and dc motors are partly replacing ac types because they are lighter, simpler, cheaper, and can use power of all frequencies.

The ac series motor is similar to the dc motor, except for a laminated field, compensating winding (larger sizes), and precautions against low power factor and poor commutation. The runaway speed is high.

In the repulsion motor, the stator is like that of

the single-phase induction motor. The rotor is similar to a dc armature. The brushes are short-circuited and set at a small angle so that stator and rotor axes differ by some 10 to 25 electrical degrees. Once this type was frequently used as a starting arrangement for single-phase induction motors. A centrifugal arrangement short-circuited the commutator and sometimes lifted the brushes at about $\frac{1}{2}$ speed. Now they are mostly superseded by capacitor motors. There is no electrical connection between stator or line and rotor.

The stator mmf has a component (field axis) which sets up the field flux, and another at right angles to it in the armature (or transformer) axis. The latter sets up the armature torque current by transformer action. When the armature turns, its conductors cut the field axis flux and generate an armature speed emf which sets up a second magnetizing component of armature current which induces the machine cemf in the transformer axis of the stator. It has an elliptical revolving field which tends to become circular at synchronism. Runaway speed is low, often not greatly above synchronism. Speed control is accomplished by adjusting voltage or series resistor or by shifting brushes. This last also permits reversing the direction of rotation.

Probable Future Trends. These include stronger magnetic materials, conducting materials with lower resistivity, and insulating materials which are usable at higher temperatures.

A. F. PUCHSTEIN

References

dc Machinery

- Gray, Alexander, and Wallace, G. A., "Principles and Practices of Electrical Engineering," New York, McGraw-Hill Book Co., 1962.
Magnussen, Carl Edward, "Direct Currents," First edition, New York, McGraw-Hill Book Co., 1929.

ac Machinery

- Arnold, E., and LaCour, J. L., "Die Synchronen Wechselstrommaschinen," Vol. 4 of "Die Wechselstromtechnik," Berlin, J. Springer, 1913.
Langsdorf, Alexander S., "Theory of Alternating Current Machinery," Second edition, New York, McGraw-Hill, Book Co., 1955.
Lawrence, Ralph R., "Principles of Alternating-Current Machinery," Third edition, New York and London, McGraw-Hill Book Co., 1940.
Richter, R., "Electrische Maschinen," Vol. 2, Berlin, J. Springer, 1930 and 1953.
Tarboux, Joseph G., "Alternating-Current Machinery," Scranton, Pennsylvania, International Text-book Co., 1947.

ac Commutator Motor

- Arnold, E., and LaCour, J. L., "Die Wechselstromkommutatormaschinen" Vol. 5, Part 2, of "Die Wechselstromtechnik," Berlin, J. Springer, 1912.

Oliver, C. W., "The A-C Commutator Motor," New York, Dr. Van Nostrand and Co., 1928.

Punga, F., "Das Funken von Kommutatormotoren," Hannover, Gebrüder, Jänecke, 1905.

Cross-references: ALTERNATING CURRENTS, ELECTRICITY, INDUCED ELECTROMOTIVE FORCE, INDUCTANCE, MAGNETISM, TRANSFORMER.

MUSICAL SOUND

Musical sound may be characterized as an aural sensation caused by the rapid periodic motion of a sonorous body while noise is that due to non-periodic motions. The above statement, originally made by Helmholtz, may be modified slightly so that the frequencies of vibration of the body fall into the limits of hearing: 20 to 20,000 cps. This definition is not clear cut; there are some noises in the note of a harp (the twang) as well as a recognizable note in the squeak of a shoe. In other cases it is even more difficult to make a distinction between music and noise. In some modern "electronic music" hisses and thumps are considered a part of the music. White noise is a complex sound whose frequency components are so closely spaced and so numerous that the sound ceases to have pitch. The loudness of these components is approximately the same over the whole audible range, and the noise has a hissing sound. Pink noise has its lower frequency components relatively louder than the high frequency components.

The attributes of musical sound and their subjective correlates are described briefly. The number of cycles per second, frequency, is a physical entity and may be measured objectively. Pitch, however, is a psychological phenomenon and needs a human subject to perceive it. In general, as the frequency of a sonorous body is raised, the pitch is higher. However, pitch and frequency do not bear a simple linear relationship. To show the relationship, a pitch scale can be constructed so that one note can be judged to be two times the pitch of another and so on. The unit of pitch is called the mel, and a pitch of 1000 mels is arbitrarily assigned to a frequency of 1000 cps. In general, it is observed that the pitch is slightly less than the frequency at frequencies higher than 1000 cycles, and slightly more than the frequency at frequencies less than 1000 cps. Pitch also depends on loudness. For a 200 cycle tone if the loudness is increased the pitch decreases, and the same happens for frequencies up to 1000 cps. Between 1000 and 3000 cps pitch is relatively independent of loudness, while above 4000 cps, increasing the loudness raises the pitch. A rapid variation in pitch when the variation occurs at the rate of from two to five times per second is called vibrato. The pitch variation in mels may be large or small but the rate at which the pitch is varied is rarely greater than five times per second. Violinists produce vibrato by sliding their fingers back and forth a minute distance on a stopped string. A variation in loudness occurring at the rate of two to five times a second is called tremolo. Singers

often produce a combination of tremolo and vibrato to give added color to their renditions.

Like frequency, intensity is a physical entity defined as the amount of sound energy passing through unit area per second in a direction perpendicular to the area. It is proportional to the square of the sound pressure, the latter being the rms pressure over and above the constant mean atmospheric pressure. Since sound pressure is proportional to the amplitude of a longitudinal sound wave (see WAVE MOTION) and to the frequency of the wave, intensity is proportional to the square of the amplitude and the square of the frequency. Sound intensity is measured in watts per second per square centimeter and, since the ear is so sensitive to sound, a more usual unit is microwatt per second per square centimeter. By way of example, a soft speaking voice produces an intensity of .1 microwatt/cm² sec, while fifteen hundred bass voices singing fortissimo at a distance 1 cm away produce 40 watts/cm² sec. Because of such large ranges of intensities, the decibel scale of intensity is normally used to designate intensity levels. An arbitrary level of 10⁻¹⁶ watts/cm² sec is taken as a standard for comparison at 1000 cps. This is very close to the threshold of audibility. At this frequency, other sound levels are compared by forming the logarithm of the ratio of the desired sound to this arbitrary one. Thus $\log I/10^{-16}$ is the number of bels a sound of intensity I has, compared to this level. Since this unit is inconveniently large, it has been subdivided into the decibel one-tenth its size; $10 \log I/10^{-16}$ equals the number of decibels (db) the sound has. A few intensity decibel levels are listed:

	db
Quiet whisper	10
Ordinary conversation	60
Noisy factory	90
Thunder (loud)	110
Pain threshold	120

While intensity levels can be measured physically, loudness levels are subjective and need human subjects for their evaluation. The unit of loudness is the phon, and an arbitrary level of zero phons is the loudness of a 1000-cps note which has an intensity level of 0 db. Sounds of equal loudness, however, do not have the same intensity levels for different frequencies. From a series of experiments involving human subjects, Fletcher and Munson in 1933 constructed a set of equal loudness contours for different frequencies of pure tones. These show that for quiet sounds (a level of 5 phons) the intensity level at 1000 cycles is about 5 db lower than an equally loud sound at 2000 cycles, for 30 cycles about 70 db lower, and at 10,000 cycles about 20 db lower. In general, as the intensity level increases, loudness levels tend to be more alike at all frequencies. This means that as a sound gets less intense at all frequencies, the ear tends to hear the higher and lower portions of sound less loudly than the middle portions. Some high fidelity systems incorporate circuitry that automatically boost the

high and low frequencies as the intensity level of the sound is decreased. This control is usually designated a loudness control.

That entity which enables a person to recognize the difference between equally loud tones of the same pitch coming from different musical instruments is called timbre, quality, or tone color. A simple fundamental law in acoustics states that the ear recognizes only those sounds due to simple harmonic motions (see VIBRATION) as pure tones. A tuning fork of frequency f , when struck, causes the air to vibrate in a manner which is very nearly simple harmonic. The sound that is heard does, in fact, give the impression that it is simple and produces a pure tone of a single pitch. If one now strikes simultaneously a series of tuning forks having frequencies f (the fundamental), $2f$, $3f$, $4f$, $5f$, etc. (overtones), the pitch heard is the same as that of the single fork of frequency f except that the sound has a different quality. The quality of the sound of the series can be changed by altering the loudness of the individual forks from zero loudness to any given loudness. Another way to alter the tone quality is to vary the time it takes for a composite sound to grow and to decay. A slow growth of an envelope even though it contains the same frequencies makes for a different tone quality than one which has a rapid growth. The difference in quality between a b-flat saxophone and an oboe is almost entirely due to the difference in growth or decay time.

A fundamental theorem discovered by the mathematician Fourier states that any complicated periodic vibration may be analyzed into a set of components which has simple harmonic vibrations of single frequencies. If this method of analysis is applied to the composite tones of musical instruments, it is seen that these tones consist of a fundamental plus a series of overtones, the intensity of the overtones being different for instruments of differing timbre. Rise and decay

times will also differ. The reverse of analysis is the synthesis of a musical sound. Helmholtz was able to synthesize sound by combining sets of oscillating tuning forks of various loudness to produce a single composite steady tone of a definite timbre. Modern synthesizers are more sophisticated. Electrical oscillators of the simple harmonic variety are combined electrically and then these electrical composite envelopes are electronically modified to produce differing rise and decay times. A transducer changes the electrical composite envelope into an acoustical one so that a sound of any desired timbre rise and decay time can be produced. An alternate way to produce similar effects is to use an oscillation known as the square wave. When this is analyzed by the method of Fourier, it is shown to consist of a fundamental plus the odd harmonics or overtones. Another kind of oscillator, a sawtooth wave, when analyzed, is shown to consist of the fundamental and all harmonics—even and odd. A square wave or a sawtooth wave produced by an appropriate electrical oscillator can be passed through an electrical filter which can attenuate any range of frequencies of the original wave. This altered wave can later be transformed into the corresponding sound wave. In this way sounds having a desired rise and decay time, plus the required fundamental and overtone structure, can be made as desired.

JESS J. JOSEPHS

References

- Helmholtz, H., "On the Sensations of Tone," New York, Dover Publications, 1954.
 Rayleigh, J. W. S., "The Theory of Sound," New York, Dover Publications, 1945.
 Josephs, J. J., "The Physics of Musical Sound," Princeton, N.J., D. Van Nostrand Co., in press.

N

NEUTRINO

The neutrino is an elementary particle postulated by W. Pauli¹ in 1930 to explain the apparent non-conservation of energy and momentum in that class of nuclear radioactivity known as beta decay. A quantitative theory of beta decay incorporating the neutrino hypothesis was formulated by E. Fermi² in 1933 in analogy with the quantum theory of radiation and served to predict the nature of the neutrino and its extremely weak interaction with matter. According to the Pauli-Fermi ideas, the neutrino (in Italian, "little neutral one") is a particle of vanishingly small and possibly zero rest mass, no electrical charge, with spin 1/2, and the ability to carry energy and linear and angular momentum. Its interaction with matter is so weak that a 3-MeV antineutrino is predicted to be capable of penetrating an astronomical thickness of matter, e.g., 100 light-years of liquid hydrogen. In 1956, a group of Los Alamos physicists³ succeeded in making a direct observation of the neutrino, $\bar{\nu}_e$, emitted from beta-decaying fission fragments produced in a powerful reactor at the Savannah River Plant operated by the du Pont Company for the U.S. Atomic Energy Commission. These investigators used giant liquid scintillation detectors to observe the inverse beta decay reaction

$$\bar{\nu}_e + p \rightarrow e^+ + n$$

where p is the target proton and e^+ and n are the product positron and neutron. The experiment consisted of observing the distinctive delayed coincidence between the prompt annihilation of the positron and the capture of the neutron by a cadmium isotope dissolved in the scintillator. In 1957, following the "overthrow" of parity conservation in weak interactions as a result of the work of Lee and Yang⁴, the character of the neutrino was further elucidated. Two kinds of neutrinos were accepted: the neutrino, ν_e , produced in beta decay in association with positrons, and the antineutrino, $\bar{\nu}_e$, produced in beta decay in association with negative electrons. The neutrino emerged as completely polarized with the spin angular momentum parallel (antiparallel) to the linear momentum for the antineutrino (neutrino). A theory of weak interactions encompassing the neutrino, in which the relativistically invariant forms, vector and axial vector,

were found to be sufficient to account for most of the known characteristics of the weak interactions, was then formulated by Marshak and Sudarshan⁵, and Feynman and Gell-Mann⁶.

In 1962, an experiment at the Brookhaven National Laboratory by a Columbia-Brookhaven group⁷, using a heavily shielded 10-ton spark chamber array, showed that the neutrino most frequently associated with the decay of the π meson differed from the neutrino produced in nuclear beta decay, thus enlarging the class of neutrinos to four: ν_e , $\bar{\nu}_e$, ν_μ , $\bar{\nu}_\mu$. It now appears that any decay or inverse process involving an electron has associated with it an electron type neutrino while any such process involving a mu meson occurs in association with a mu-type neutrino. It was suggested at the 1964 International High Energy Conference in Dubna that these neutrinos be distinguished by the names muneutrino and elneutrino.

A search with high-energy machines⁸ for a particle which may be responsible for the weak interaction and which decays into a muneutrino and a muon or an elneutrino and an electron shows that, if it exists, it is more massive than 1.8 protons. Neutrino physics now encompasses low-energy reactions using fission reactors, work in the structure of the weak interaction using giant electronuclear machines, and the beginnings of various studies of natural neutrino sources such as the sun and high-energy cosmic rays.⁹

F. REINES

References

1. Pauli, W., Jr., "Rapports Septuems Conseil Physique, Solvay, Bruxelles, 1933," Paris, Gautier-Villars, 1934.
2. Fermi, E., *Z. Physik*, **88**, 161 (1934).
3. Reines, F., and Cowan, C. L., Jr., *Phys. Rev.*, **92**, 830 (1953); Cowan, C. L., Jr., Reines, F., Harrison, F. B., Kruse, H. W., and McGuire, A. D., *Science*, **124**, 103 (1956); *Phys. Rev.*, **117**, 159 (1960).
4. Lee, T. D., and Yang, C. N., *Phys. Rev.*, **105**, 1671 (1957).
5. Marshak, R. E., and Sudarshan, E. C. G., *Phys. Rev.*, **109**, 1860 (1958); proceedings of Padua-Venice Conference on Mesons and Newly Discovered Particles, Italy, September 1957.
6. Feynman, R. P., and Gell-Mann, M., *Phys. Rev.*, **109**, 193 (1958).

7. Danby, G., Gaillard, J. M., Goulianos, K., Lederman, L. M., Mistry, N., Schwartz, M., and Steinberger, J., *Phys. Rev. Letters*, 9, 36 (1962).
8. Reported by the CERN Laboratory Group at the Dubna conference.
9. Further references and a more extensive discussion of the neutrino are given in Reines, F., *Science*, 141, 778 (1963).

Cross-references: CONSERVATION LAWS AND SYMMETRY, ELEMENTARY PARTICLES, RADIOACTIVITY, WEAK INTERACTIONS.

NEUTRON

Discovery. The discovery of the neutron by Chadwick¹ in 1932 represented a great step forward in the investigation of nuclei of atoms. Chadwick found that a radiation emitted when α -rays from polonium reacted with beryllium could project protons from a thin sheet of paraffin wax. Although the radiation itself produced no observable ionization when passing through a gas, the protons released from the paraffin were detected in an ionization chamber. Inability to produce ionization was interpreted as a lack of electric charge. From measurements of the ionization from the protons, Chadwick deduced that the so-called beryllium radiation must consist of neutral particles with a mass very nearly equal to that of the proton. He announced the discovery of the neutron, a previously unknown particle. It has been confirmed that the neutron has no charge and a mass of 1.008982 atomic mass units. Thus it is heavier than the proton by 0.00139 mass unit. The introduction of the neutron into nuclear structure produced a sharp change in the previously held ideas. Lacking knowledge of the neutron, masses of atomic nuclei had been attributed solely to protons. The number of protons required on this basis for most nuclei greatly exceeded the known charge number. In an attempt to solve this dilemma, a number of electrons were assigned to each nucleus to adjust the charge number to the proper value. This compromise created an even greater problem, that of accommodating so many electrons in the small space occupied by a nucleus. Bringing the neutron into the picture, it is now only necessary that a nucleus contain protons to equal the charge number with the rest of the mass contributed by neutrons. No electrons are required.

Detection. Because it is a neutral particle the neutron can be detected only by means of a secondary charged particle which it releases in passing through matter or by means of the radioactivity which the neutron can induce in stable elements. Protons may be projected by collisions with neutrons in hydrogenous material and the ionization from the protons can be measured in an ionization chamber, as in the original experiment with neutrons. Secondary charged particles may be the direct result of nuclear disintegration produced by neutrons, as in the case of the reaction $B^{10} + n^1 \rightarrow Li^7 + \alpha$.

Commonly the radioactivity induced in stable elements by neutron capture serves to detect neutrons, and this technique is known as the activated foil method. Also fission may be utilized for detection of neutrons by placing fissionable material inside an ionization chamber and observing the ionization generated by the fission fragments.

Decay. The neutron in the free state undergoes radioactive decay. Elaborate experiments by Robson² were required to identify the products of the decay and to measure the half-life of the neutron. He showed that the neutron emits a β -particle and becomes a proton. The half-life was found to be 12.8 minutes. In stable nuclei, neutrons are stable. In radioactive nuclei, decaying by β -emission, the neutrons decay with a half-life characteristic of the nuclei of which they are a part.

Energies. The kinetic energy of neutrons has an important bearing on the behavior of neutrons when interacting with nuclei. These kinetic energies may range from near zero to as much as 50 MeV. It is, therefore, natural to classify neutrons in terms of energy according to their properties in each range of energy. For example, energies from zero to about 1000 eV are usually called slow neutrons. Because they are more readily captured by nuclei than faster neutrons, slow neutrons are responsible for a large number of nuclear transformations. When slow neutrons have velocities in equilibrium with the velocities of thermal agitation of the molecules of the medium in which they are situated, they are called thermal neutrons. The distribution of these velocities approaches the Maxwell distribution

$$dn(v) = Av^2 e^{-\frac{Mv^2}{2kT}} dv$$

where v is the neutron velocity, M its mass, k is Boltzmann's constant and T the absolute temperature. In the slow neutron range of energies, various atomic nuclei show strong absorption (capture) of neutrons at fairly well-defined energies. Neutrons having energies corresponding to those of the absorption bands are called resonance neutrons. Frequently, neutrons with energies greater than 1000 eV and less than 0.5 MeV are termed intermediate neutrons. In more general terms, all neutrons with energies greater than 0.5 MeV are called fast neutrons. The practical upper limit of neutron energy is set by the devices so far developed for accelerating charged particles to extremely high energies.

Magnetic Moment and Spin. Alvarez and Bloch³ succeeded in measuring the moment of the magnetic dipole associated with the known spin of 1/2 possessed by the neutron. More refined measurements by Cohen, Corngold, and Ramsey⁴ of the magnetic moment μ_n yielded a value of

$$\mu_n = -1.913148 \text{ nuclear magnetons}$$

Interactions with Nuclei. Neutrons may be scattered or captured by heavy nuclei. Scattering

may be elastic, resulting only in the change of direction of the neutrons, or inelastic in which the neutron loses part of its energy to the scattering nucleus. Collisions with light nuclei, in absence of capture, result in communicating considerable fractions of the neutron energy to the target nucleus. A neutron colliding head-on with a proton will give practically all its kinetic energy to the proton. As the mass of the target nucleus increases, the transfer of energy decreases, in accordance with the laws of conservation of energy and momentum. The loss of energy by mechanical impact is utilized in slowing down fast neutrons, a process known as moderation. Slow neutrons are more useful, for example, in the production of radioelements from stable elements by neutron capture. A good moderator should have low mass and a small capture cross section. The rate r of capture of neutrons from a neutron flux F (neutrons $\text{cm}^{-2} \text{sec}^{-1}$) incident on a layer of matter having N nuclei per square centimeter is given by

$$r = F\sigma N$$

where σ is the complete probability of capture. Replacing r by dN/dt and writing the flux as nv , where n is the number and v is the velocity of the neutrons, we have

$$\frac{dN}{dt} = -nv\sigma N$$

which integrated gives

$$N = N_0 e^{-nv\sigma t}$$

where N is the number of unchanged nuclei in the target area at time t and N_0 is this number at time $t = 0$. The cross section σ is so named because it has the dimensions of an area. The unit for the cross section is the barn equal to 10^{-24}cm^2 . When, as is often the case, σ is proportional to $1/v$, the advantage of slow neutrons in capture interactions becomes apparent. When the value of σ departs sharply from that predicted by the $1/v$ law, it usually increases over a narrow range of energies, and we have what is called a RESONANCE. Slow neutron cross sections are customarily quoted for thermal neutrons at 20°C, corresponding to a value of v of 2200 m/sec. In Table I values of σ are given for a few representative stable elements.

Additional interactions of neutrons with nuclei include the release of charged particles by neutron-induced nuclear disintegration. Commonly known reactions are $n-p$, $n-d$, and $n-\alpha$. In these cases, the incident neutrons may contribute part of their kinetic energy to the target nucleus to effect the disintegration. Hence, more than mere neutron capture is involved. Then there is usually a lower threshold for the neutron energy below which the reaction fails to occur. Another important reaction involving neutrons is fission which may occur under different conditions for either slow or fast neutrons with appropriate fissionable material.

TABLE I. THERMAL NEUTRON CAPTURE CROSS SECTIONS*
 $v = 2200 \text{m/sec}$

Element	σ (barns)
Boron	759
Cobalt	38
Cadmium	2 450
Gadolinium	46 000
Gold	99.8
Helium	0
Lead	0.170
Oxygen	<0.0002

* See reference 5 for data for other elements.

Sources of Neutrons. Any nuclear reaction in which neutrons are released might serve as a source of neutrons. In the initial experiments on neutrons, an $\alpha-n$ reaction was used. Because of the charge on the α -particle, it must have a high kinetic energy to penetrate a nucleus. Thus polonium α -particles could release neutrons from beryllium. Such a natural source produces relatively few neutrons. The yield of neutrons from charged particle reactions can be increased manyfold by the use of particle accelerators. Here large numbers of charged particles of high energy can be used in the bombardment of the target to release numerous neutrons. Frequently deuterons or protons are used for the bombardment. A far more prolific source is the nuclear reactor. Fission of uranium is usually the source of the neutrons in this case. A nuclear reactor, as usually constructed, generates neutrons of different energies in various parts of its structure. Neutrons of suitable energy for a given experiment may be brought outside the reactor through channels into appropriate sections of the reactor.

Structure of the Neutron. Ordinarily the neutron is regarded simply as a particle which is a component of nuclei and which can exist only briefly in the free state. For many purposes this view is sufficient. However, it has become obvious from experiments, for example, in very high-energy accelerators, that the neutron must have a complex structure. This view is reinforced by the nature of the decay of the neutron. A β -particle is ejected from the neutron on decay, but it is quite certain that the electron did not exist within the neutron prior to the decay. Rearrangements of an internal structure of the neutron must provide the energy for the formation and ejection of the β -particle. One theory would have the neutron consist of a proton and a π^- meson bound together so that they oscillate between a completely bound state and a more loosely bound state. Such a theory might also explain the feeble interaction which has been observed between electrons and neutrons at very short range. At present, it may be sufficient to say that the neutron must have a complex

internal structure of a nature at present not very clearly understood.

Additional information on the neutron may be obtained from the books listed below.^{6,7}

L. F. CURTISS

References

1. Chadwick, J., *Proc. Roy. Soc. London Ser. A*, **136**, 692 (1932).
2. Robson, J. M., *Phys. Rev.* **83**, 349 (1951).
3. Alvarez, L. W., and Bloch, F., *Phys. Rev.*, **57**, 111 (1940).
4. Cohen, V. W., Corngold, N. R., and Ramsey, N. F., *Phys. Rev.*, **104**, 283 (1956).
5. "American Institute of Physics Handbook," Second edition, Section 8, New York, McGraw-Hill Book Co., 1963.
6. Curtiss, L. F., "Introduction to Neutron Physics," Princeton, N.J., D. Van Nostrand, 1958.
7. Evans, Robley D., "The Atomic Nucleus," New York, McGraw-Hill Book Co., 1955.

Cross-references: COLLISIONS OF PARTICLES, CROSS SECTIONS AND STOPPING POWER, ELECTRON, FISSION, NUCLEAR REACTIONS, NUCLEAR REACTORS, PROTON, RADIOACTIVITY, RESONANCE.

NEUTRON ACTIVATION ANALYSIS

Neutron activation analysis is a method of elemental analysis based upon the quantitative detection of radioactive species produced in samples via nuclear reactions resulting from neutron bombardment of the samples.

Types of Neutron Reactions. The neutron-induced reactions are of two main types: (1) those induced by very slow (thermal) neutrons, having energies of about 0.025 eV, and (2) those induced by fast neutrons, those having energies in the range of MeV.

All stable isotopes are capable of capturing thermal neutrons, but with characteristic reaction cross sections which vary widely from isotope to isotope, even of the same element. Promptly following the capture of a thermal neutron by a stable nucleus, the compound nucleus de-excites itself by the emission of one or more "prompt" gamma-ray photons. If the resulting product nucleus is a radioisotope, its later decay can be detected and can be of use in the activation analysis detection of that element. Thermal-neutron capture reactions are therefore referred to as " (n, γ) " reactions. For example, in the determination of vanadium, with thermal neutrons, some of the V^{51} stable nuclei present in the sample to be analyzed undergo the $V^{51}(n, \gamma)V^{52}$ reaction. Vanadium-52 is radioactive, decaying with a half-life of 3.77 minutes, emitting a β^- particle and a 1.43-MeV gamma-ray photon.

Fast neutrons predominantly interact with nuclei by means of (n, p) , (n, α) , $(n, 2n)$, and (n, n') reactions. Whereas the thermal-neutron capture reaction forms an isotope of the original element, but now one mass unit higher, the $(n, 2n)$ fast-neutron reaction forms an isotope of the same

element one mass unit lower. An example of this type of reaction is the $N^{14}(n, 2n)N^{13}$ reaction. Nitrogen-14 is the abundant stable isotope of nitrogen, whereas N^{13} is radioactive, decaying with a half-life of 10.0 minutes by positron emission. The (n, n') fast-neutron type of reaction, termed a "neutron inelastic scattering" reaction, forms an excited state (nuclear isomer) of the original nucleus, with unchanged mass number (A), but a measurable half-life. The $Se^{77}(n, n')Se^{77m}$ reaction is a good example of this type of reaction. Selenium-77 is one of the stable isotopes of selenium, whereas Se^{77m} is a radioactive isomer of Se^{77} that decays with a half-life of 17.5 seconds, emitting an isomeric-transition 0.16 MeV gamma-ray photon. In (n, p) reactions, the product nucleus has the same mass number as the original nucleus, but is a different element, namely, lower by one unit in atomic number (Z). A widely utilized reaction of this type is the $O^{16}(n, p)N^{16}$ reaction. Oxygen-16 is the principal stable isotope of oxygen; N^{16} is a radioactive isotope of nitrogen, decaying with a half-life of 7.35 seconds, emitting exceptionally high-energy beta particles and gamma-ray photons. In (n, α) reactions, the product nucleus has a mass number 3 units lower than the original nucleus, and a Z that is 2 units lower than originally. For example, in the fast-neutron detection of phosphorus, the $P^{31}(n, \alpha)Al^{28}$ reaction is often utilized. Normal phosphorus consists entirely of P^{31} ; Al^{28} is a radioactive isotope of aluminum, decaying with a half-life of 2.27 minutes, emitting a β^- particle and a 1.78-MeV gamma-ray photon in each disintegration.

Theory of the Method. When a sample containing N nuclei of a given type (a particular Z and A) is exposed to a flux ϕ of neutrons, resulting in a particular nuclear reaction having a cross section σ , the rate of formation of product nuclei by this reaction, in nuclei per second, is simply $N\phi\sigma$. The units of ϕ and σ are neutrons per square centimeter per second and square centimeters per nucleus, respectively. If the product nucleus is a radioactive species, some of these nuclei will be decaying while the irradiation is going on. If the irradiation of the sample with neutrons is continued for a long time, compared with the half-life of the radioisotope formed, a steady state, or "saturation" condition will be reached, in which previously formed nuclei are decaying at the same rate that new ones are being formed, thus with no further increase in the disintegration rate of this particular product with continued irradiation at that flux. Therefore, the saturation activity of a given species, at zero decay time (i.e., just at the conclusion of the irradiation), is $A_0(\text{satn}) = N\phi\sigma$.

At intermediate irradiation times (t_i), the activity of a particular induced species (A_0 , expressed in disintegrations per second, i.e., dps) is equal to $N\phi\sigma S$, where S is a "saturation" term that is equal to $1 - e^{-0.693t_i/t_{0.5}}$, where $t_{0.5}$ is the half-life of the radioactive species. The saturation term is dimensionless and ranges only from 0 (at $t_i = 0$) to 1 (at $t_i = \infty$). It rapidly approaches

a value of one, asymptotically, acquiring values of 1/2, 3/4, 7/8, 15/16, . . . at $t_1/t_{0.5}$ values of 1, 2, 3, 4, Because of this rapid approach of S to its maximum value, it is pointless to activate a sample for a period of time longer than a few half-lives of the radioactive species of interest. Longer irradiation merely generates more interfering activities of longer half-lives.

In the basic activation equation, σ is the isotopic cross section for the particular type of nuclear reaction, for neutrons of a specified energy. It assumes that the neutron flux, and energy, are constant throughout the sample. The N term is itself equal to $wfaN_A/AW$, in which w is the weight of the sample (in grams), f is the weight fraction of the element in the sample, a is the fractional abundance of the target stable isotope among all the stable isotopes of the element, N_A is Avogadro's number, and AW is the ordinary chemical atomic weight of the element. In actual analyses, of course, either f or the product wf is the unknown quantity.

Neutron Sources. In neutron activation analysis work, the most widely used neutron sources are (1) research-type nuclear reactors, and (2) small accelerators. Modern research reactors are mostly of the pool type and operate at power levels of 10 to 1000 kW, providing thermal-neutron fluxes of 10^{11} to 10^{13} n cm⁻² sec⁻¹ and fission spectrum fast-neutron fluxes of about the same magnitude. The small accelerators used in neutron activation analysis work are largely low-voltage (100 to 200 kV) Cockcroft-Walton deuteron accelerators, capable of producing up to about 10^{11} 14-MeV neutrons/sec from a tritium target (with a 1-mA deuteron beam current), via the $H^3(d,n)He^4$ reaction. Samples of typical size (~ 1 cm³) can thus be exposed to a 14-MeV neutron flux of about 10^9 n cm⁻² sec⁻¹. In a moderator, these can produce a thermal-neutron flux of the order of 10^8 n cm⁻² sec⁻¹. Unfortunately, at full-power operation, the lifetime of the tritium target is only of the order of an hour to a few hours. Some work is carried out also with lower-energy neutrons generated by the $Be^9(d,n)Be^{10}$ reaction with a 2-MeV positive-ion Van de Graaff accelerator, or by the $Be^9(x,n)Be^8$ reaction with bremsstrahlung produced by a 3-MeV electron Van de Graaff accelerator. These produce thermal-neutron fluxes in the range of 10^8 to 10^9 n cm⁻² sec⁻¹. Isotopic sources, such as Po^{210} -Be, Pu^{239} -Be, and Am^{241} -Be, also generate neutrons, but the maximum thermal-neutron flux attainable with such sources is only about 10^5 n cm⁻² sec⁻¹. They are useful for teaching purposes, but not for real analytical work.

Sensitivities for Various Elements. As the available neutron flux increases, the level of induced activity per unit mass of an element also increases; hence, the sensitivity of detection is higher, i.e., the limit of detection is lower. With a nuclear-reactor thermal-neutron flux of 10^{13} n cm⁻² sec⁻¹, a maximum t_1 of one hour, and reasonable counting efficiencies, it is found that the median limit of detection for some 70 elements is about 10^{-3} μ g. A few of these elements can be detected

down as low as 10^{-7} μ g; a few only to about 10 μ g. The method, with such high neutron fluxes is the most sensitive known method for over half the elements of the periodic system. With 1-gram samples, the μ g absolute sensitivities correspond to parts-per-million (ppm) concentration sensitivities. Samples ranging from minute samples up to 10 grams or somewhat more can be irradiated and analyzed. With longer irradiation periods, the detection limits for about half of the 70 elements (those forming longer-lived induced activities) can be reduced further. The most sensitively detected elements (limits $\leq 10^{-3}$ μ g) are the following: Ag, As, Au, Br, Co, Cu, Dy, Er, Eu, Ga, Ge, Ho, I, In, Ir, La, Lu, Mn, Na, Nb, Pd, Pr, Re, Rh, Sb, Sm, Sr, U, V, W, and Yb. Some elements are more sensitively determined by activation with fast neutrons than with thermal neutrons.

At the lower (10^8 to 10^9) thermal-neutron fluxes attainable with the small accelerators, the limits of detection are of course 10^4 to 10^5 times higher than for the reactor 10^{13} flux. At a 10^9 thermal-neutron flux, the sensitivities for the same 70 elements thus range from 10^{-3} μ g to 0.1 gram, with a median of 10 μ g. With a 14-MeV neutron flux of 10^9 n cm⁻² sec⁻¹, a number of additional elements can be detected fairly sensitively, down to levels of 0.1 to 10 μ g, e.g., N, O, F, Si, P, Cr, and Fe.

Forms of the Method. In practical analytical work, one does not employ the basic activation analysis equation ($A_0 = N\phi\sigma S$), per se, but instead uses a comparator technique. When samples are to be analyzed for one or more elements, standard samples of these elements are activated at the same time as the unknowns and then are counted in an identical manner (counting efficiency ϵ). When the counting data are corrected to the same decay time, then, A (unknown)/ A (standard) grams element in unknown/grams element in standard, since ϵ , f , AW , ϕ , σ , and S are the same for both unknown and standard. In the equation, A refers to the counting rate (rather than disintegration rate) of the radioisotope formed by the element in question—in unknown and standard, respectively. Not only is the comparison technique simpler, but it removes any dependence upon literature values of σ , and experimental values of ϵ and ϕ , which often are not accurately known. At levels well above the limits of detection, careful application of this comparison technique results in precisions and absolute accuracies in the range of ± 1 to 3 per cent of the value.

The method is employed in two different forms: the purely instrumental form and the radiochemical separation form. The instrumental form is fast and nondestructive, and is based upon the quantitative detection of induced gamma-ray emitters by means of multichannel gamma-ray spectrometry. It is the preferred method where it applies. Induced activities are identified by the energies of their gamma-ray photopeaks observed in the NaI(Tl) scintillation counter pulse-height spectrum of the activated samples. The amount

of the element present in a sample is computed usually from the photopeak (total absorption peak) height or area of its gamma ray, or one of its principal gamma rays, compared with that of the standard.

Where interferences from other induced activities are very serious, and cannot be removed adequately by decay, spectrum subtraction, or computer solution, one must turn to the radiochemical separation form of the method. Here the activated sample is put into solution and equilibrated chemically with measured amounts (typically 10 mg) of added carrier of each of the elements of interest, before chemical separations are carried out. The element to be detected needs then to be recovered in chemically, and radiochemically, pure form, but it need not be quantitatively recovered, since the carrier recovery is measured and the counting data are then normalized to 100 per cent recovery. This form of the method is slower, but it applies to pure beta emitters, as well as to gamma emitters, and it does eliminate interfering activities. It is free of the usual complications of microconcentration analysis: high blanks from reagent impurity, and losses by adsorption and coprecipitation.

Neutron activation analysis is now a well-established method of elemental analysis, carried out in many laboratories and utilized by many more through available commercial activation analysis services. It is now widely applied in almost every branch of science, engineering, and medicine, where either its great sensitivity (at high fluxes) or its speed (with the instrumental form of the method), or both, are used to advantage.

VINCENT P. GUINN

References

- Bowen, H., and Gibbons, D., "Radioactivation Analysis," London, Oxford University Press, 1963.
- Lyon, W. S., "Guide to Activation Analysis," Princeton, N.J., D. Van Nostrand, 1964.
- Koch, R. C., "Activation Analysis Handbook," New York, Academic Press, 1960.
- Albert, P., "L'analyse par Radioactivation," Paris, Gauthier-Villars & Cie, 1960.
- Texas A & M University, "Proceedings of the December 1961 International Conference on Modern Trends in Activation Analysis," Texas A & M University, College Station, Texas, 1961.
- Guinn, V. P., and Steele, E. L., "Neutron Generators and Their Uses," *Natl. Acad. Sci., Natl. Res. Council, Nucl. Sci. Ser.* (1965).
- Meinke, W. W., in "Chemistry Research and Chemical Techniques Based on Research Reactors," pp. 17-59, 73-82, 95-114, International Atomic Energy Agency, Vienna, 1963.
- Anders, O. U., "Gamma-Ray Spectra of Neutron-Activated Elements," Dow Chemical Company, Midland, Michigan, 1964.
- Guinn, V. P., and Schmitt, R. A., "Determination of Pesticide Residues," in Gunther, F. A., Ed., "Residue Reviews," Vol. 5, pp. 148-174, Berlin, Springer, 1964.
- Guinn, V. P., "Non-Biological Applications of Neutron Activation Analysis in Forensic Studies," in Curry, A. S., Ed., "Methods in Forensic Science," Vol. 3, pp. 47-68, London, Interscience Publishers, 1964.

Cross-references: ACCELERATORS, PARTICLE; ISOTOPES; NEUTRON; NUCLEAR REACTIONS; NUCLEAR REACTORS; RADIOACTIVITY.

NEUTRON DIFFRACTION

An experiment by Laue, Friedrich and Knipping in 1912 demonstrated that x-rays were a form of electromagnetic radiation, with a wavelength of the same order of magnitude as the distance apart (10^{-8} cm) of atoms in crystals. This meant that beams of x-rays could be diffracted by crystals in a rather similar way to that in which an optical diffraction grating, in which the elements are separated by about 10^{-3} cm, will produce a spectrum for visible light. As a result of Laue's discovery, a technique for studying the underlying structure of solids by "x-ray diffraction" has grown up. For any given solid, the end product of this technique is a specification of the shape and content of the building block, or "unit cell," out of which the solid is built. The content is specified in terms of "electron density," and it follows that the various atoms or ions which make up the molecule of the substance can be identified.

A rather similar, but in some respects a much more powerful, technique has grown up using beams of neutrons instead of x-rays. A neutron is often thought of simply as a particle, with a mass approximately equal to that of a hydrogen atom, but in terms of wave mechanics a beam of neutrons can be regarded as a wave motion. If the neutrons are moving with velocity v , then they can be considered to have a wavelength equal to h/mv , where m is the mass of the neutron and h is Planck's constant. If such a neutron beam is scattered by a solid, it will be distributed in space as if it were radiation of this wavelength. It so happens that for neutrons having energies equivalent to a temperature of a few hundred degrees centigrade, which are readily obtainable from nuclear reactors, the wavelength is about 10^{-8} cm, i.e., 1 \AA , which, as we have seen above, is about equal to the interatomic distance in solids. It was shown in 1936 that neutrons, then obtainable only from a radium-beryllium source, could indeed be diffracted by solids. However, it is only since nuclear reactors have produced intense beams of suitable neutrons that the application of diffraction techniques to the study of solids has proved worthwhile.

Since high-intensity neutron beams are only available at a limited number of research institutions throughout the world, we shall be concerned only with their application to problems which cannot be solved by any other method. In particular, we shall enquire what can be achieved with neutrons which cannot be found out by

using a beam of x-rays, and we shall see the answers to this question by making a comparison of the ways in which atoms and solids scatter x-rays and neutrons. *X-rays* are scattered by the outer, extranuclear, electrons in an atom, and it is for this reason that x-ray diffraction studies produce a picture of electron density. It follows that heavy atoms, such as lead and uranium which contain many electrons, will predominate in these pictures and that a one-electron atom, i.e., hydrogen, can be located and detailed with much less accuracy. On the other hand, *neutrons* are scattered not by electrons but by the nucleus of an atom, and the way in which the scattering power increases with the mass of the atom is very far from being a steadily increasing function. The scattering power or, more precisely, what we call the "scattering length" arises from the summation of two quite separate effects. The first of these depends on the size of the nucleus, which has a radius proportional to the cube-root of the atomic weight, so that this effect does indeed increase with atomic weight, but nevertheless fairly slowly. Superimposed on this scattering, however, is resonance scattering, which depends in a complicated way on the actual structure of the nucleus and on its energy levels. This additional scattering often varies quite considerably from atom to atom, and sometimes from isotope to isotope, as we advance up the periodic table. When we combine together the two effects, and thus assess the resultant scattering by a nucleus, we find that it varies quite irregularly from atom to atom and this is illustrated for elements at the lower end of the periodic

system in Fig. 1. It will, however, be noted that there is a relatively small spread of values among these scattering lengths. The mean value for all the nuclei which have so far been measured is 0.62×10^{-12} cm, and practically all elements have values which lie between a half and twice this average. As a result of this we find that most elements are roughly equally "visible" to neutrons, though there are a few very interesting exceptions. The practical outcome of this is that hydrogen atoms can be located quite accurately in whatever environment they are found, and this has meant important advances in our knowledge of the role of hydrogen bonds and molecules of water of crystallization in building up the structures of both inorganic and organic crystals. At the same time, we have often been able to get much improved information on the thermal motion of molecules, particularly in those common cases where hydrogen atoms are found on the outside of molecules and which, therefore, provide a very good index of the molecular movement. The technique of detection becomes much more powerful if we can use *deuterated* material, instead of ordinary hydrogen. Deuterium has a neutron scattering length of 0.65×10^{-12} cm and is, therefore, a "good average" atom, whereas ordinary hydrogen, at 0.38×10^{-12} cm, is somewhat below average. This comparison provides a very good example of a difference between the scattering behavior of two different isotopes of an element, arising from differences in the nuclear structures.

Another important field of chemistry to which neutron diffraction has contributed some useful

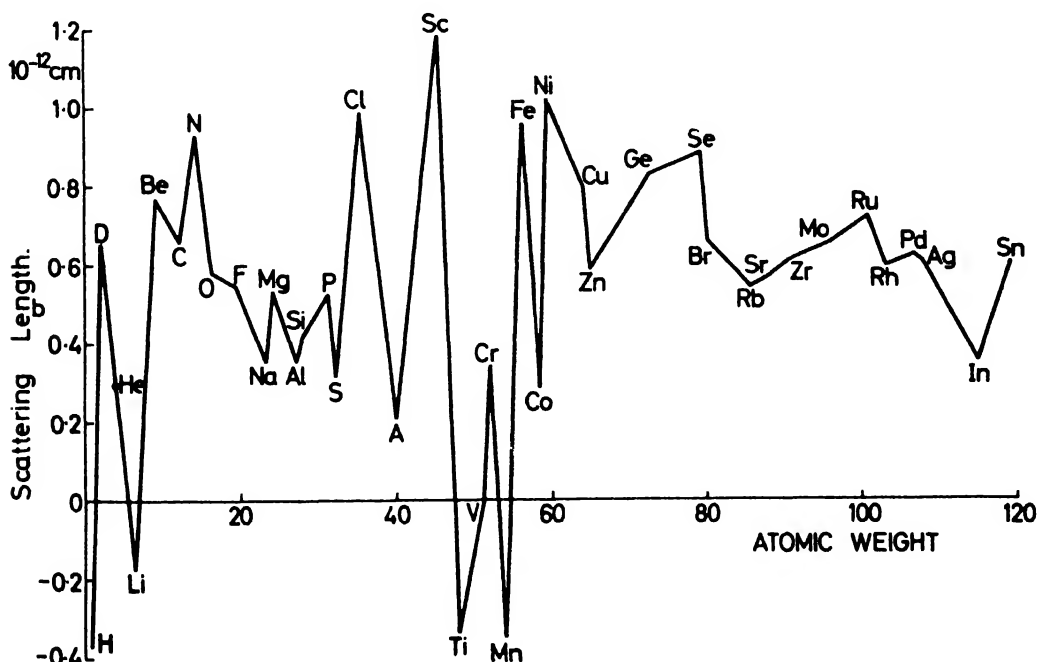


FIG. 1. The variation of the nuclear scattering amplitude of elements for neutrons, shown as a function of atomic weight, in units of 10^{-12} cm.

results is in the study of the compounds of uranium, and post-uranic elements, with nitrogen and oxygen. In the case of x-ray studies, the 92 electrons of a uranium atom completely overshadow the seven and eight of nitrogen and oxygen respectively. For neutrons, however, the value of the scattering length for uranium (0.85×10^{-12} cm) is actually less than the value for nitrogen (0.94×10^{-12} cm) and is only fractionally greater than that of oxygen (0.58×10^{-12} cm).

So far we have been considering the process whereby the neutron is scattered by the atomic nucleus, and this is a process which occurs for all atoms. There is, however, an additional scattering which takes place for *magnetic* atoms, i.e., for atoms which have a resultant magnetic moment on account of the fact that the atoms contain unpaired electrons. Examples of this are an atom of iron in metallic iron, which appears to contain 2.2 unpaired electrons, and the doubly-charged manganese ion Mn^{++} , which contains 5 unpaired electrons, in manganese salts. Such atoms or ions scatter additional neutrons, making an additional contribution to the scattering length by an amount which is proportional to the magnetic moment. If the magnetic moments in such a material are not arranged in any regular single direction, but point haphazardly as in a *paramagnetic* material, then there will not be any well-defined diffracted beams but there will be a broadly distributed contribution to the scattered background. This contribution may be a little difficult to identify but, nevertheless, the identification can be achieved and the phenomenon can be confirmed. In other magnetic materials, however, all the magnetic moments in a single domain lie parallel to a single direction, and in the particular case of a *ferromagnetic* material they all point *algebraically* in the same sense. In this circumstance the magnetically scattered neutrons contribute specifically to the diffracted beams and the intensity of these is observed to vary with increase of temperature, falling to a minimum at the approach of the Curie temperature, above which no ferromagnetic alignment takes place. In the case of antiferromagnetic materials, in which the moments lie parallel to a single direction but alternately up and down with opposite algebraic sense, the neutron data are extremely informative. In such a material it will be appreciated that, from a *magnetic* point of view, the repeat distance (considering the alternate $+$ and $-$ moments) is twice the repeat distance which is apparent when only the *chemical* nature of the atoms is considered. This means that extra diffraction spectra will be produced at smaller angles of scattering, corresponding to what would happen if the inter-line spacing of an optical diffraction grating was doubled. The existence of antiferromagnetism can, therefore, be detected very directly by noting the appearance of these extra spectra, particularly if the neutron diffraction pattern is compared with either an x-ray pattern or with a neutron pattern taken at a higher temperature at which the regular magnetic

arrangement has broken down. Such a comparison of results obtained at two different temperatures is illustrated in Fig. 2. Results such as these have established the antiferromagnetic structures of a variety of materials and have demonstrated the true nature of ferrimagnetism, as for example in the ferrites in which moments are directed in both positive and negative directions but with a net balance in one direction.

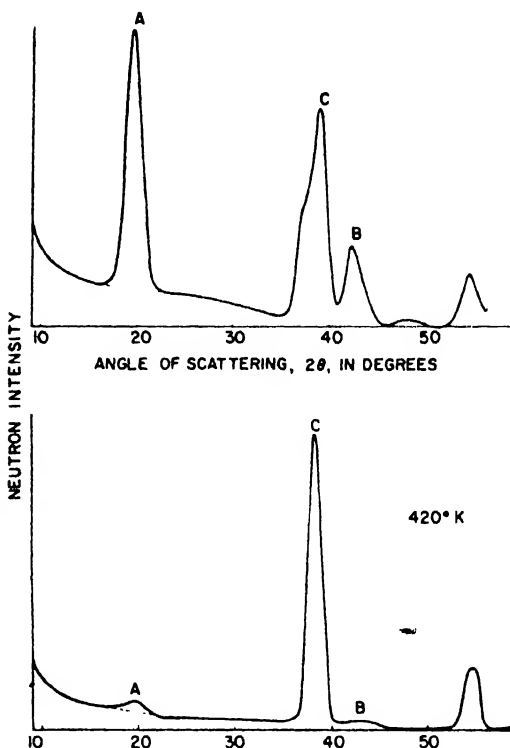


FIG. 2. A comparison of neutron diffraction patterns taken at 4 and 420 K for an antiferromagnetic alloy, $Au_2Mn_{17}Al_{13}$. At the lower temperature, intense magnetic lines A and B appear, but at the higher temperature, where very little magnetic order remains, these extra lines have practically disappeared. The patterns also show the composite nature of the nuclear scattering line C at low temperature, which occurs because the crystal symmetry changes from cubic to tetragonal when the magnetic order becomes established.

Moreover, with further research, it has been demonstrated that these structures are only the simplest examples of a wide range of "magnetic architecture" which it is now possible to draw in detail as a result of study with neutron beams. This later work, devoted to the iron group of transition elements and the elements of the rare earth group, has identified a variety of non-collinear arrangements of magnetic moments such as the spiral spins in $MnAu_2$, the umbrella structure in CrSe, composite structures in holmium and erbium and the complicated structures,

not yet fully understood, which occur in metallic chromium.

In our discussion so far, we have implicitly assumed that when a neutron is scattered by an atom it is scattered *elastically* and does not lose any of its energy. This is no more than a first approximation to the truth, because atoms are by no means rigidly fixed but are in vibration about their mean positions because of their possession of thermal energy. A neutron which makes collision with an atom may, therefore, lose or gain a quantity of energy. If we keep in mind the fact that atoms in a solid are not isolated, and that the movement of one atom will invariably affect to some degree the motion of its neighbors, it becomes fruitful to regard the interchange of energy as occurring between the neutron and the lattice vibrations of the solid. Indeed we speak of "phonons" which are the embodiment in *particle* form, from the point of view of wave mechanics, of the quanta of energy among the crystal vibrations. If we could measure accurately the interchanges of energy, then we could learn about, and indeed study in detail, the phonon spectra and the dispersion law for the solid. In fact, for *neutron* scattering, but not for x-rays, such a measurement can be made and this gives a quite unique value to the use of inelastic neutron scattering for studying solids. The particular supremacy of neutrons becomes clear if we consider the actual energies of a neutron and x-ray quantum which possess the same wavelength. We find in fact that the latter is roughly 10^5 times larger. Thus, whereas the energy of a neutron of wavelength 1 \AA is about equal to that of a quantum of crystal energy yet the energy of a 1 \AA x-ray is 10^5 times greater. It follows, therefore, that if a neutron gains or loses such a quantum, then its own energy will be greatly changed; for example, it could easily be roughly doubled or halved. On the other hand such an interchange for an x-ray would be quite insignificant and the resulting change of wavelength could not be detected. It is in fact possible, therefore, to measure both momentum and energy changes in neutron-phonon interchanges, and this information leads directly to the details of the dispersion law in the solid.

As the intensity of the beams of neutrons available from research reactors steadily increases, with roughly a ten fold increase each ten years, these techniques become progressively more powerful and determinative, leading to a steadily widening view of solids and liquids in the several unique respects which we have discussed. The limitations in the use of these techniques are those set by the limited availability of suitable nuclear reactors, and it may be fairly said that many promising applications have not yet been tested.

G. E. BACON

References

Bacon, G. E., "Neutron Diffraction," Second edition, London, Oxford University Press, 1962.

Bacon, G. E., "Applications of Neutron Diffraction in Chemistry," Oxford, Pergamon Press, 1963.

Ringo, G. R., in "Handbuch der Physik," Vol. 32, pp. 552-642, Berlin, Springer-Verlag, 1956.

Shull, C. G., and Wollan, E. O., "Solid State Physics," Vol. 2, pp. 138-217, New York, Academic Press Inc., 1956.

I.A.E.A., "Inelastic Scattering of Neutrons in Solids and Liquids" proceedings of a symposium in Vienna (1960), 1961.

Cross-references: DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS, MAGNETISM, NEUTRON, PARAMAGNETISM, X-RAYS.

NOISE, ACOUSTICAL

Strictly defined, noise is any unwanted sound, whether pleasant or unpleasant. More commonly, however, sounds that are unpleasant and disturbing, or that mask desired sound, are termed noise. Thus, noise, in a general sense, may be thought of as any sonic disturbance. Depending upon the degree of pitch distribution, intensity, and persistence, noise can range from being merely annoying, to hazardous or injurious. In our highly industrialized society, with its rapid growth of energy-producing and converting systems, noise has become a major problem. Some of its harmful effects are interference with mental and skilled work, impairment of sleep, creation of emotional disturbances, damage to hearing and a deterioration of health and well-being. Consequently, the control and reduction of noise has become an important science.

Adequate measuring means are a prime requirement in the scientific control and reduction of noise. Even then, the problem of establishing a true relationship between the subjective and objective properties of noise is difficult because of the many different aspects of human reaction to noise. The first relationship between the subjective and objective measurement of sound is the simplified rule relating *loudness in sones* to the *loudness level in phons*. The loudness scale in sones is proportional to the average person's estimate of the loudness. Thus, the loudness level, P , of a given sound, in phons, is numerically equal to the median sound pressure level, p_L , of a free progressive wave at a frequency of 1000 cps presented to listeners facing the source, which is judged by the listeners to be equally loud. The sound pressure level, p_L , is defined as

$$p_L = 20 \log_{10} \frac{p}{p_0} \text{ (decibels)} \quad (1)$$

where p is measured sound pressure, in microbars, and $p_0 = 2 \cdot 10^{-4}$ microbars.

The relation between sones, S , and phons, P , is given by

$$S = 2^{(P-40)/10} \quad (2)$$

Referring to Eq. (2), a loudness level of 40 phons produces a loudness of one sone, and a loudness level of 80 phons produces a loudness level of sixteen sones, etc.

The simplest means for the measurement of noise is the sound level meter, an instrument comprising a microphone, an amplifier, frequency weighting networks, and an output meter. The characteristics of the frequency weighting networks in the meter are based upon the equal loudness contours of hearing for different levels.

More sophisticated means for measuring noise include octave band and one-third octave band sound analyzers that supply information on the sound level in various frequency ranges. These analyzers are used for research on the reduction of machine noise, transmission and other areas where information on the sound levels in specific frequency bands is required. Narrow-band analyzers may be used to obtain the spectrum of a noise. The sound-pressure spectrum level is that level within a frequency band of 1 cps. This level is plotted against frequency to obtain the spectrum frequency characteristic of the noise. If the spectrum level of a noise is known, sophisticated means may be used to relate the objective to the subjective qualities of the noise.

As noted earlier, noise abatement has become an important science. For instance intensive research on the quieting of automobiles has been in progress for three decades, with outstanding results; some of the major problems remaining to be solved involve wind- and road-induced noise. Similarly, research has been carried out on the reduction of noise of all types of household appliances employing motors, fans, compressors,

pumps, gears, and other moving parts. Another phase of acoustical engineering involves methods for reducing the transmission of sound through the walls, floors, ceiling and partitions in all manner of buildings or enclosure by the use of construction and materials based on fundamental acoustical principles.

The noise in typical environments and noise produced by various sources are given in Table 1.

The masking effect produced by noise reduces the intelligibility of speech. For example, if the speaker and listener are separated by 5 feet, the levels of noise that will barely permit reliable word intelligibility are 51 db for normal conversation, 57 db for raised speech, 63 db for very loud speech, and 69 db for shouting.

A person subjected to high noise levels for long periods of time may suffer considerable impairment of hearing. The use of ear protectors may provide sufficient insulation under such conditions.

The establishment of any scale of noisiness judgment is difficult because human response and reaction are very complicated. However, for random noise a relationship between the objective and subjective responses may be outlined in a general form as shown in Table 2.

TABLE 2. RELATION BETWEEN OBJECTIVE AND SUBJECTIVE RESPONSES TO NOISE

Sound Pressure Level (db)	Noise Evaluation
40-50	Quiet
70-80	Moderate
80-90	Noisy
90-100	Very noisy

HARRY F. OLSON

References

Harris, C. M., "Handbook of Noise Control," New York, McGraw-Hill Book Company, 1957.

Cross-references: ACOUSTICS; ARCHITECTURAL ACOUSTICS; HEARING; MEASUREMENTS, PRINCIPLES OF; MUSICAL SOUND.

NUCLEAR INSTRUMENTS

Nuclear instruments are those devices which are used to make the group of measurements which is commonly described as nuclear radiation detection. These measurements are understood to encompass not only the indication of the presence of nuclear radiation but also the determination of the amount, energy, and related properties. The nuclear instrument may be considered to consist of two parts, a detector and a measuring apparatus. The interaction of the radiation with the instrument takes place in the detector, while the measuring apparatus takes the output of the

TABLE 1. NOISE LEVELS FOR VARIOUS SOURCES AND LOCATIONS*

Source or Description of Noise	Noise Level (db)
Threshold of pain	130
Hammer blows on steel plate 2 ft	114
Riveter 35 ft	97
Factory	78
Busy street traffic	68
Large office	65
Ordinary conversation 3 ft	65
Large store	63
Factory office	63
Medium store	62
Restaurant	60
Residential street	58
Medium office	58
Garage	55
Small store	52
Theatre	42
Hotel	42
Apartment	42
House, large city	40
House, country	30
Average whisper 4 ft	20
Quiet whisper 5 ft	10
Rustle of leaves in gentle breeze	10
Threshold of hearing	0

* Olson, Harry F., "Acoustical Engineering," p. 256, Princeton, N.J., D. Van Nostrand Co., 1957.

detector and performs the required analysis to accomplish the measurement.

Nuclear instruments are either pulse type or non-pulse type. In the pulse-type instrument, the output of the detector is a series of signals separated or resolved in time, each signal representing the interaction of a nuclear particle with the detector. If the pulse feature of the detector output is used, e.g., in counting particles or in measurement of the distribution in energy of particles the nuclear instrument is said to be pulse type.

In the non-pulse type of operation of a nuclear system, no attempt is made to resolve the signals of individual particles. Instead, the average effect due to many interactions of the radiation with the detector is measured directly. This type of instrument, which can be described as a mean-level type detection system, is typified by the current-type ionization system. The current output is proportional to the number of particles incident upon the detector per unit time.

For the study of the physics of the interaction of NUCLEAR RADIATION with the detectors, it is convenient to divide the numerous forms of radiation into classes, based on the mode of interaction with matter as in a detector, and to discuss the properties of a prototype of each class. The radiations which are convenient as prototypes are protons, electrons, fission fragments, gamma rays, and neutrons.

The prototype protons, fission products, electrons, and all other radiations consisting of charged particles each interact with matter primarily by the production of excitation and ionization of the matter through which it is passing, with ionization playing the major role. The nuclear particles lose energy at a rate w per ion pair formed. Values of w for particles passing through gases range from 25 to 50 eV per ion pair. The three types of particles differ primarily in their penetrating ability; for example, a 0.03-MeV electron will pass through a 2.2 mg/cm² of air, while a proton would need 0.9 MeV of energy and a fission product 64 MeV to penetrate the same thickness.

Gamma rays and neutrons, being uncharged radiations, differ in their primary mode of interaction from the types discussed above. However, both gamma rays and neutrons undergo primary interactions with matter which produce secondary charged particles that interact with the detector, thereby producing effects which make possible the detection as in the case of the radiation types which consist of charged particles. The interaction of gamma rays with matter produces secondary electrons through the principal mechanisms of photoelectric effect (see PHOTO-ELECTRICITY), Compton scattering and pair production, as well as through several other secondary mechanisms.

In the case of neutrons, there are also several mechanisms by which the interaction with matter takes place, each of which is the basis for a potential method of detection. The most useful ones are: (1) neutron-induced transmutations in

which the product particles make detection possible; examples are (n, α) , (n, p) , (n, γ) , and $(n, \text{fission})$ reactions; (2) neutron-induced transmutations, leading to radioactive nuclei, the subsequent decay of which makes the detection possible; (3) elastic scattering of neutrons (for example, by a proton) in which the recoil proton can be detected.

Several nuclear radiation detectors depend for their operation on the IONIZATION that is produced in them by the passage of charged particles. This group of detectors includes ionization chambers, proportional counters, Geiger-Müller counters, semiconductor-radiation detectors, cloud chambers, and spark chambers. For the uncharged particles, such as neutrons and gamma rays, the charged particles which are required for the production of ionization originate by the secondary processes referred to above.

In other detectors, excitation and sometimes molecular dissociation also play important roles. These phenomena, in combination with ionization, bring about the LUMINESCENCE in scintillation detectors and the latent images in photographic emulsions. Also, molecular dissociation is important in chemical detection systems, i.e., those devices that function through the occurrence of certain chemical reactions.

One of the oldest, but still most widely used, types of detector employs a gas-filled chamber. Depending on the mode of operation of this chamber, the detector type is known either as an ionization chamber, a proportional counter, or a Geiger-Müller (G-M) tube. The principles of operation and the differences in the three modes of operation can be explained by the use of Fig. 1 and 2. The system shown in Fig. 1 consists of a gas-filled chamber containing a central electrode that is well insulated from the chamber wall. A potential V is applied between the central electrode and the chamber wall through the high resistance R shunted by the capacitor C_2 .

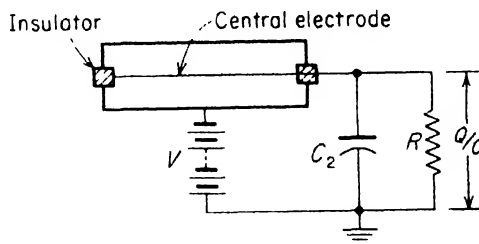


FIG. 1. Schematic diagram for pulse operation of gas-filled chamber. (Reprinted from Price, W. J., "Nuclear Radiation Detection," p. 42, New York, McGraw-Hill Book Company, Inc., 1964).

Assume that the passage of a nuclear particle releases N_1 ion pairs within the chamber. The positive and negative charges within the chamber move toward the chamber wall and central electrode, respectively, because of the direction of the electric field. Under the condition that

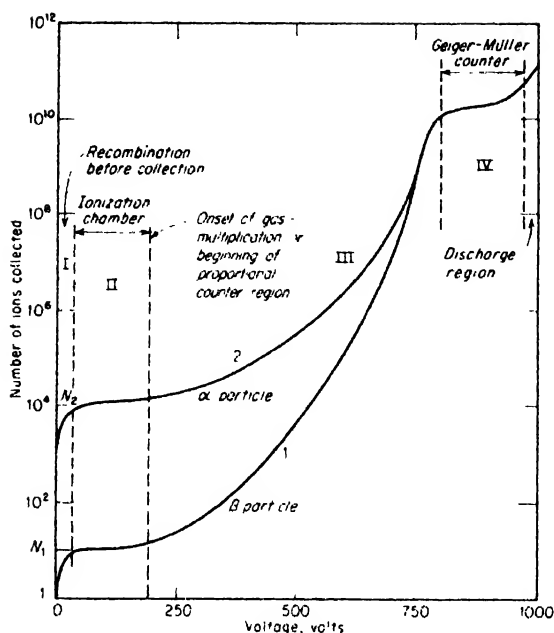


Fig. 2. Pulse-height vs applied-voltage curves to illustrate ionization, proportional, and Geiger-Müller regions of operation. [Reprinted from Montgomery, C. G., and Montgomery, D. D., *J. Franklin Inst.*, 231, 449 (1941)]

the time constant RC_2 is much greater than the time required for the collection of the charge, the charge Q appearing on the capacitor per particle as a function of V is given by curve 1 in Fig. 2. For a particle producing a larger number of ion pairs N_2 , curve 2 is obtained.

These curves can be divided into four main regions. In region I there is a competition between the loss of ion pairs by recombination and the removal of charge by collection on the electrodes. With increasing field, the drift velocity of the ions increases; therefore the time available for recombination decreases, and the fraction of the charge which is collected becomes larger.

In region II the recombination loss is negligible, and the charge collected is

$$Q_1 = N_1 e \text{ and } Q_2 = N_2 e$$

The change in voltage across the capacitor C_2 is

$$\Delta V_1 = \frac{N_1 e}{C} \quad \text{and} \quad \Delta V_2 = \frac{N_2 e}{C}$$

where C is the sum of the ionization-chamber capacity and C_2 . This region is referred to as the saturation region or the ionization-chamber region.

In region III the collected charge is increased by a factor M through the phenomenon of gas multiplication. The electrons which are released in the primary ionization are accelerated sufficiently to produce additional ionization and thus

add to the collected charge. At the onset of region III, the multiplication M for a given applied voltage is independent of the initial ionization, thus preserving the proportionality of pulse sizes. This strict proportionality breaks down with increase in applied voltage until, at the upper limit of region III, the pulse size is independent of the initial ionization. This region, in which gas multiplication is employed while at the same time a dependence of the collected charge on the initial ionization remains, is known commonly as the proportional region. The upper end of it is designated as the region of limited proportionality.

In region IV the charge collected is independent of the ionization initiating it. Rather, gas multiplication increases the charge to a value that is limited by the characteristics of the chamber and the external circuit. This region is known as the G-M region.

Ionization chambers find wide-scale use in non-pulse-type applications, for example, in monitoring of radiation for personnel protection. Pulse-type operation is not attractive in the ionization chamber region because of the relatively small-size pulse released by an individual nuclear particle. On the other hand, in the proportional region, gas multiplication increases the size of pulses while at the same time information concerning the size of the initial ionization is maintained. Therefore, proportional counters are useful as counting devices (with the ability to discriminate between particles making different amounts of ionization) and for energy measurements. Geiger-Müller counters are widely used for counting electrons between particles and gamma rays inasmuch as the large discharge pulse is triggered by the very small ionization which these particles produce. However, they have no capability for differentiating between radiation types or for measuring the energy of the particles which trigger the discharge.

Inasmuch as Lord Rutherford utilized light scintillations in his famous alpha-scattering experiments, the scintillation detector is one of the oldest methods of nuclear radiation detection. However, the modern scintillation detector, a device of great utility, only evolved after the development of special photomultiplier tubes for the purpose. Figure 3 shows the schematic of a modern scintillation detector for counting. When charged particles pass through certain substances, ionized and/or excited states are produced, which, during their return to the normal states, produce light flashes, or scintillations. By coupling the scintillators with a photomultiplier tube, a pulse of charge can be passed to an electronic system, as shown in Fig. 3, thus making counting possible. Also, if the electronic counter is replaced by electronic equipment for measuring the pulse size, the scintillation detector is very useful for energy measurements.

Various scintillators are available including crystals of inorganic and organic materials, liquids, powders, plastics, glasses, and gases. Therefore, the designer has available a wide range

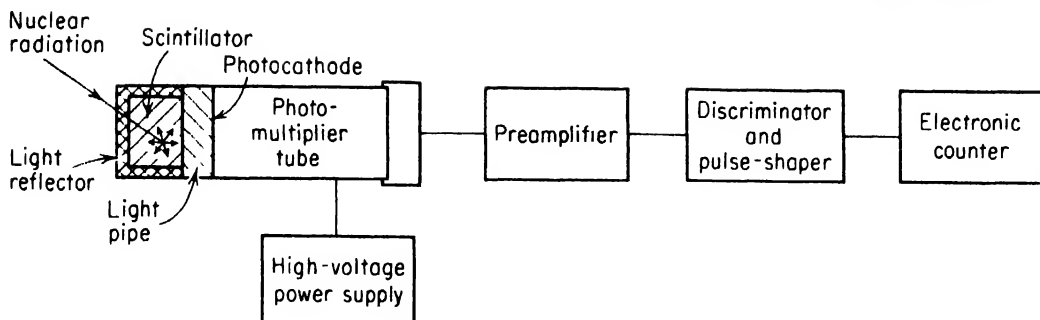


FIG. 3. Schematic diagram of a scintillation detector (Reprinted from Price, W. J., "Nuclear Radiation Detection, p. 160, New York, McGraw-Hill Book Company, Inc., 1964)

of sizes, shapes, and compositions, and consequently can optimize conditions for a wide variety of measurement problems including a wide range of radiation types and energies. In addition, the time duration of the pulse produced can be much shorter than those in the gas-filled chambers, and the density of the detection medium is usually much higher. As a result of these several advantages, there is a large group of applications in which the scintillation detectors have superiority over gas-filled chambers as well as over other detectors. Gamma-ray spectroscopy is an outstanding example of an important class of measurement made possible by the scintillation detectors.

The newest of the important nuclear radiation detectors is commonly referred to as the semiconductor radiation detector, after the fact that it is the properties of semiconductor materials that make these detectors possible. In this detector an electric field is set up across a semiconducting medium of low electrical conductivity. Usually the low conductivity region is the charge-depletion region in a semiconductor diode operated at reverse bias. When a charged particle passes through the semiconductor, electron-hole pairs are produced in it. The charges are caused to separate by the electric field, and the resulting electrical signal can be transmitted to the electronic measuring system to give important information concerning the particle detected.

The semiconductor radiation detector has several key properties, the combination of which accounts for the rapid rise of importance of this detector type. These properties include linearity of pulse height vs energy, very rapid response time, high resolution, convenient dimensions, thin windows, insensitivity to magnetic fields, variable sensitivity with respect to particle energy, relative insensitivity to gammas and neutrons, and relatively low cost. For example, the availability of semiconductor radiation detectors has revolutionized the spectrometry of charged particles.

Several other types of detectors are in use and are quite important in specialized measurements. These other detectors include nuclear-track plates, cloud chambers, spark chambers, Cerenkov counters, chemical detectors, calorimetric

methods and various special types of neutron detectors. For details on the principles of operations and the characteristics of these detectors, as well as more information on the three detector types described above, the reader is referred to the textbook "Nuclear Radiation Detection", by W. J. Price (New York, McGraw-Hill Book Co., Inc., 1964) and the references included in that book.

WILLIAM J. PRICE

Cross-references: IONIZATION; LUMINESCENCE; MEASUREMENTS, PRINCIPLES OF; NUCLEAR RADIATION; RADIATION, IONIZING, BASIC INTERACTIONS; RADIOACTIVITY.

NUCLEAR RADIATION

Nuclear radiation results from the transitions of atomic nuclei. The two chief types of transition in natural radioactivity are those in which the number of constituent particles of a given nucleus (nuclide) are changed by the emission of one or more particles, and those in which there is a rearrangement of the particles of a given nuclide, without change in number, such that the nuclide passes from a state of higher energy to a state of lower energy.

Soon after the discovery of radioactivity in 1896, it was found that naturally radioactive substances emit three kinds of radiation: alpha (α), beta (β) and gamma (γ) rays. The first two consist of high-speed charged particles, alpha and beta particles respectively. They are called *particle* rays to distinguish them from the gamma rays discovered by Villard in 1900. The latter were shown by a series of experiments to be a form of high-frequency electromagnetic radiation traveling with the speed of light. However, in discussing the nature of these rays, it must be remembered that experiments clearly show that the so-called *particles* may exhibit distinctly *wave-like* characteristics, and conversely the higher the frequency of the gamma radiation the more pronounced becomes the *particle-like* character of its individual quanta or photons. From such considerations springs present-day quantum mechanical theory.

Two of the most important differences between alpha rays and beta rays are: (1) they are deflected in opposite directions by a magnetic field indicating that they are oppositely charged, and (2) the alpha particle is far more massive than the beta particle. Early measurements indicated that alpha particles may be emitted with speeds up to $1/15$ the speed c of light, and beta rays with speeds up to $0.96c$. The rest masses of the two particles and their energy equivalents in millions of electron volts (MeV) are

$$M_{\alpha}(\text{alpha particle}) = 6.645 \times 10^{-27} \text{ kg} = \\ = 3727.2 \text{ MeV}$$

$$m_{\beta}(\text{beta particle}) = 9.109 \times 10^{-31} \text{ kg} = \\ = 0.511 \text{ MeV}$$

Thus the alpha particle is nearly 7300 times more massive than the beta particle. The reason for such a difference in the masses was found early in this century. In experiments begun in 1903, Rutherford showed that an alpha particle is the doubly charged (positive) nucleus of a helium atom. Experiments of Becquerel and others identified the beta particle as the then recently discovered, negatively charged electron.

With the discovery of what has been called "artificial" or induced radioactivity by Irene Curie and her husband Frederick Joliot in 1933, it was found that positive electron (positron) emission may occur in nuclides whose instability results from the nucleus possessing "too much charge for its mass." This is, of course, only another form of beta emission represented as β^+ whereas ordinary electron emission is represented as β^- .

Certain characteristics of nuclear radiation such as their great penetrating power, their capacity to ionize a gas, affect a photographic plate, or produce serious or fatal burns, and the ability of alpha rays to produce scintillations in a fluorescent screen were early recognized. The exact penetrating power of a particular ray depends upon the energy with which it is emitted by the parent nucleus, but in general it may be said that alpha rays may usually be stopped by a few sheets of paper and beta rays by a few millimeters of aluminum. Alpha and beta rays, because of their particle character, have a more sharply defined range than gamma rays, and a significant fraction of the gamma radiation may pass through a number of centimeters of metal shield. This is because particles lose speed and kinetic energy in both ionizing and non-ionizing collisions with other particles, whereas a gamma-ray photon always travels at the same speed, and the beam of photons is usually weakened by some process such as scattering or absorption of photons. The frequencies of photons may also be reduced by the so-called *Compton collisions* (see COMPTON EFFECT).

One of the most important characteristics of alpha and beta rays is their capacity to produce ions in a gas and render it conducting. In normal air, an alpha ray will produce between 20,000

and 70,000 ion pairs per centimeter of path, depending on the speed of the ray, with maximum ionization being reached near the end of the path. A beta ray on the other hand may only produce 200 ion pairs or less per centimeter. Thus such rays lose energy rapidly in passing through a gas, and rays of a particular type and energy have a rather sharply defined range (see IONIZATION).

The range of alpha particles in air at 76 cm mercury pressure (15°C) varies from 2.7 cm for alpha particles from uranium to 8.62 cm for alpha particles from thorium C' (Po^{212}). The former are emitted with energies of 4.2 MeV while the latter, the most energetic of any from a naturally radioactive substance, are emitted with energies of 8.6 MeV.

In 1929, Rosenblum discovered the fine structure of alpha rays. That is, alpha rays from a single type of nuclide may not all have exactly the same energy and range but often consist of two or more groups with slightly different but sharply defined ranges and consequently different initial energies. This led to the ultimate recognition of the existence of different energy levels in the nucleus and is one of the most important properties of alpha rays in sharp contrast with the nature and behavior of beta rays. Along with the discovery of nuclear energy levels came recognition of the origin of gamma radiation. After the emission of an alpha (or a beta) particle, if the nucleus is left in an excited state of higher energy than the lowest or ground state, gamma radiation occurs as the result of a transition to a lower-energy state.

Beta rays have ranges in air at 76 cm mercury pressure (15°C) varying from a fraction of a centimeter for those of a few thousand volts energy to more than 400 cm for those of energies of 1 MeV. However, the beta rays from a single type of nuclide have a wide distribution of ranges and energies, the distribution being approximately continuous up to a certain maximum for the particular type of nuclide. The lack of a characteristic disintegration energy and range for the various beta particles from a given type of nuclide results from the nature of the emission process as first explained by Fermi in 1934. When a beta particle is emitted, it is necessary that at least one other particle (a NEUTRINO) also be emitted in order to conserve angular momentum of spin. Conservation of energy then demands that any energy this neutrino possesses must subtract from the energy that might have been carried off by the beta particle in the process of emission from some sharply defined energy level in the nucleus. Only the maximum possible or "end-point" energy of a given beta-ray emission curve is therefore significant for that particular nuclide. On the other hand, gamma rays result from the return of a nuclide from a higher-energy state of excitation to a lower state, or to the ground state, and have sharply defined values characteristic of the states of that nuclide. Different groups of beta rays having different end-point energies may originate from the same type of nuclide and represent different nuclear energy levels.

The experimental detection of the neutrino (and antineutrino) in the years following 1956 confirms the main aspects of the Fermi theory, and adds another particle and antiparticle, albeit of zero mass and zero charge, to those types which may be ejected from the atomic nucleus. *Nuclear radiation* may involve all of these types of rays and also the neutrons and other particles resulting from occasional spontaneous fission of a nucleus or from the intense fission in a nuclear reactor.

In addition to natural and induced radioactivity and nuclear fission, many types of nuclear transformation or disintegration may occur in cosmic-ray phenomena or in bombardment of nuclei by high-energy rays. Most of the particles produced in such occurrences are themselves unstable and decay by the emission of other particles. For instance, a neutron decays to a proton by the emission of a negative electron and an anti-neutrino. A pion (π meson) may decay to a muon (μ meson) and a neutrino, and the muon may decay to an electron and two neutrinos. To complicate matters, the neutrinos and antineutrinos in the two decay processes seem to be slightly different. In a wider sense, all the nuclear particles and antiparticles emitted by any nuclear process must be included in the forms of nuclear radiation.

ROGERS D. RUSK

References

- Evans, R. D., *The Atomic Nucleus*, New York, McGraw-Hill Book Co., 1955.
 Kaplan, I., "Nuclear Physics," Second edition, Reading, Mass., Addison-Wesley Publishing Co., 1963.
 Rusk, R. D., "Introduction to Atomic and Nuclear Physics," Second edition, New York, Appleton-Century-Crofts, 1964.

Early Work

- Rutherford, E., Chadwick, J., and Ellis, C. D., "Radiations from Radioactive Substances," New York, The Macmillan Co., 1930.

Cross-references: ELECTRON; ELEMENTARY PARTICLES; FISSION; NEUTRON; NUCLEAR REACTIONS; NUCLEAR STRUCTURE; POSITRON; PROTON; RADIATION, IONIZING; BASIC INTERACTIONS; RADIOACTIVITY.

NUCLEAR REACTIONS

Several years after the discovery of the nucleus by scattering α -particles from gold, Rutherford and his collaborators noticed that if air were exposed to the flux of α -particles, occasionally a very penetrating particle was observed. After some nuclear detective work, this phenomenon was explained in the following way: the nitrogen nucleus and the α -particle react to produce an isotope of oxygen and an energetic proton. In the chemical notation, such a reaction may be written as $N^{14} + \alpha \rightarrow O^{17} + p$, where the superscripts are the atomic mass numbers of the elements in question.

Nuclear reactions may take place only when a target nucleus and a projectile come close enough together for the nuclear forces to take effect. The range of nuclear forces is very short, about 1.5 fermis (1 fermi = 10^{-13} cm). Since nuclei and all the massive projectiles except neutrons are charged positively, they repel each other, and if they are to be brought into sufficiently intimate contact to interact, the energy of the projectile must equal or surpass the repulsive electrostatic force. Typically, for protons on light nuclei, the energy required is 1 or 2 MeV, rising to 15 MeV for protons on very heavy, highly charged nuclei. Thus to produce nuclear reactions by the collision of charged particles, we must first accelerate one of them to an energy sufficient to overcome the electrostatic repulsive force. In all nuclear reactions, energies are usually given in million electron volts and masses in atomic mass units (amu). One MeV is the energy acquired by a particle of one electronic charge as it is accelerated by an electrical potential of one million volts. An atomic mass unit is currently defined as one-twelfth the mass of a neutral carbon-12 atom.

A nuclear reaction just as its chemical counterpart, may be exoergic (kinetic energy is liberated) or endoergic (kinetic energy is absorbed). For example, Rutherford's original reaction is endoergic, consuming 1.19 MeV which is converted into mass of the product particles according to Einstein's $E = mc^2$. In the case of endoergic reactions, energy must be supplied in the form of projectile kinetic energy to make the reaction possible.

In the course of a nuclear reaction, the projectile and the target may fuse completely, or parts of nuclear matter, neutrons, protons, or clusters of these, may be transferred from the target to the projectile or vice versa. The most common projectiles are ISOTOPES of hydrogen, protons or deuterons, and α -particles which are nuclei of helium atoms. Because of their low charge, these feel the lowest electrostatic repulsion from the target. Heavier projectiles are, however, frequently used, and these include lithium, carbon, nitrogen, and oxygen ions; a few experiments have been done with projectiles as heavy as Ar^{40} . Neutrons form a special class of projectile, since they are not charged and need not possess any large amount of energy to overcome a repulsive barrier. In fact, the slower the neutrons are, the more likely they are to interact with the target simply because they spend more time in the vicinity of each target nucleus. Nuclear reactions may also be initiated by PHOTONS (electromagnetic radiation quanta) of very high energy, greater than 5 to 10 MeV. The photons are produced as x-rays when high-energy electrons from a betatron, synchrotron, or linear accelerator impinge on a target. Since photons are not charged electrically, there is no barrier, but all photonuclear reactions are endoergic, that is they require several-million-electron-volt x-rays to take place.

Rutherford's original reaction would now be commonly written thus: $N^{14}(\alpha, p)O^{17}$. The target

nucleus comes first, the projectile and the emitted particle or particles appear in that order inside the parentheses, and the residual nucleus is last. The superscript gives the atomic mass of the isotope. Some common abbreviations are p for proton, n for neutron, d for deuteron (H^2), t for triton (H^3), γ for γ -ray.

The projectile energy range over which nuclear reactions are important varies from a fraction of an electron volt for neutrons to several hundred million electron volts per atomic mass unit for charged particles. At very high energies, the projectiles appear to interact with the individual neutrons and protons of the target rather than with the nucleus as a whole, and we leave the domain of nuclear reactions to enter the field of ELEMENTARY PARTICLE interactions.

Nuclear reactions are literally the foundation on which our world is built. The energy of the sun is nuclear in origin, deriving principally from the fusion of four protons into a helium nucleus, in the course of which 22.7 MeV are released in each fusion. The fusion is not a direct four-body reaction, but rather proceeds by stages through many two-body reactions (see SOLAR ENERGY SOURCES). Not only is our chief source of energy of nuclear origin, but the constituents of the earth are also the result of long-gone nuclear reactions, principally a series of (n, γ) processes which served to build up the elements in the earth as we now find them. One particularly pleasing success of nuclear reaction theory and experiment is the fact that the abundance of the elements and their isotopic ratios as they occur in nature can be calculated simply on the basis of the probability with which neutrons are captured by various nuclei, these probabilities having been measured in nuclear research centers.

Not only nature's energy source is of nuclear origin. The two mightiest sources of man-made energy, nuclear FISSION and nuclear FUSION, derive from nuclear reactions.

In fission, a neutron is captured by a uranium nucleus which splits (fissions) into, say, a Ba and Zr nucleus, in the process releasing about 200 MeV of energy and some neutrons which in turn split other U nuclei in the vicinity. This leads to the familiar chain reaction which, if controlled, is used to produce power by converting the fission energy (heat) into either motive power or electricity. If the chain reaction is allowed to proceed without control, a violent explosion results, i.e., an atomic bomb. We note parenthetically that what is commonly known as atomic energy should really be called nuclear energy since its source is not the entire atom, but only its nucleus. Burning of coal, on the other hand, is atomic energy, since heat is derived from the combination of a carbon atom with two oxygen atoms to form CO_2 .

Fusion as a source of energy derives from a reaction such as $H^2(d, n)He^3$ in which 3.3 MeV are released. This process is similar to the source of solar energy, in that hydrogen nuclei fuse to produce helium. It has not proved possible yet to control the nuclear fusion reaction in such

a way that it will proceed slowly. Many laboratories are currently working on this. An uncontrolled fusion reaction has, however, been achieved—it constitutes the energy source of the hydrogen bomb.

It is clear that reserves of coal and oil must someday be exhausted. When this happens, man will necessarily turn to nuclear reactions, fission, or fusion for his sources of power. At the present time, 1965, nuclear fission power plants account for only 0.6 per cent of the electricity generated in the U.S.; however, it is expected that 7 per cent of the power will be of nuclear origin by 1980, and 45 per cent by the year 2000.

The terrors of war and the blessings of abundant power both come from nuclear reactions. But that is not all. Perhaps the most significant contribution to all sciences has been the use of radioactive elements produced in nuclear reactions. Let us take a typical example. Consider the reaction $C^{13}(n, \gamma)C^{14}$. This reaction produces radioactive C^{14} which has a half-life of 5700 years. The radioactive carbon decays with the emission of an electron which can be counted with a suitable detector. Thus we are able to locate individual atoms of carbon and separate them from all others which are not radioactive. Such radioactive atoms are called tracers, and by using a variety of them, C^{13} , F^{18} , P^{32} , S^{35} , I^{131} , Au^{198} , all produced by some sort of nuclear reaction, unprecedented advances have been made in biology, chemistry, metallurgy, physiology, and medicine. Furthermore, radioactive isotopes in large amounts can be used instead of x-rays for treatment of cancer, for metallography, and for food preservation.

By means of nuclear reactions such as $Am^{241}(B^{11}, 4n)Fm^{248}$, scientists have been able to make new elements not found in nature. Some of these, for example plutonium and californium, have found important uses as reactor fuels or portable power sources. The preparation of new elements has played a decisive role in our understanding of the chemistry of heavy elements.

We turn now to a description of nuclear reactions themselves. These may be regarded as proceeding in two principal ways. First, the colliding particles may fuse, their components get thoroughly mixed in a very "hot" compound nucleus. This compound nucleus may exist in a heated state for a period varying between 10^{-16} and 10^{-20} second, a time we normally consider imponderably short, but which is nevertheless long on the nuclear time scale when compared with the transit time of a nucleon across the nucleus (10^{-22} second). The compound nucleus boils off fragments, mainly neutrons, protons, and α -particles, and in this way cools down to a normal energy content. In its final state, it is called a residual nucleus, and it may be radioactive or stable depending on the details of the reaction. If the target, the projectile, and the energy of the projectile are known, it is possible to predict what the residual nucleus will be and thus, if certain isotopes are wanted, one can tailor the reaction accordingly. We note here, however,

that the exact mechanism of these reactions is very complicated and is not really understood in a fundamental way.

In the other kind of reaction, the nuclei do not fuse, but only part of the nuclear matter is transferred from one nucleus to the other. This is often called a direct reaction, and includes stripping reactions, pickup reactions, and transfer reactions. A typical example is $\text{Al}^{27}(d, p)\text{Al}^{28}$, where a neutron is stripped from the deuteron and caught by the Al nucleus, while the remaining proton stays relatively undisturbed on its original course. The nuclei in such reactions do not come into intimate contact, and the reaction is fast, on the time scale of 10^{-22} second.

We recall now that there is at present no fundamental understanding of the nucleus, i.e., we do not know precisely how it is built up from its constituent neutrons and protons, and what the properties of the nuclear force are which holds it together (see NUCLEAR STRUCTURE). An important use of nuclear reactions is to throw light on this problem. By a careful study of nuclear reactions where all the details such as the precise energy of the projectile, the nature, energy, and angle of emission of the reaction products are measured, much insight has been gained concerning the nucleus itself. The problem is really twofold: we have to understand and be able to express mathematically the mechanism of a nuclear reaction, and then we must proceed to extract information about the participating nuclei from it. At present both aspects are only promising, so that our deeper understanding seems to await either a clear breakthrough or a painstaking development of more detailed experiments and ever more complicated theoretical calculations. The prize, however, is worth reaching for: its promise is to reveal the nature of nuclei, the smallest stable constituents of matter, and the laws which govern their interactions to give man immense power for war or for peace.

A. ZUCKER

References

Two advanced books which deal with nuclear reactions are:

Endt, P. M., and Demeur, M., "Nuclear Reactions," Vol. 1, Amsterdam, North-Holland Publishing Company, 1959.

Eisenbud, L., and Wigner, E. P., "Nuclear Structure," Princeton, N.J., Princeton University Press, 1958.

Cross-references: ELEMENTARY PARTICLES, FISSION, FUSION, ISOTOPES, NEUTRON, NUCLEAR STRUCTURE, PHOTON, PROTON, RADIOACTIVITY, SOLAR ENERGY SOURCES, TRANSURANIUM ELEMENTS.

NUCLEAR REACTORS

A nuclear reactor is a facility in which nuclear fuel is assembled for the purpose of supporting a sustained, controlled, neutron-fission chain reaction.

The controlled fissioning of atoms of nuclear

fuel, resulting from interaction with neutrons, is the crucial event in a nuclear reactor. All of the manifold applications of "atomic energy," from nuclear power to atomic bombs, to radioactive isotopes, depend directly on the fissioning of fuel atoms or on the by-products therefrom. Basically, all these applications of atomic energy derive from just three consequences or by-products of fission: (1) release of heat energy, (2) release of direct radiation including neutrons, and (3) generation of "fission products," the fragments of the fissioned atoms which themselves are highly radioactive isotopic species.

Role of Neutrons. Within an assemblage of nuclear fuel, neutrons are initially obtained from spontaneous fission or from an inserted neutron source.* Some of these neutrons cause fission of fuel atoms. Each fission releases, among other radiations, two or more additional neutrons which, in turn, may either collide with other fuel atoms and cause them to fission or may suffer a variety of other fates. Whether a neutron causes fission or ends its life some other way depends on the quantity and geometrical distribution of fuel, the moderating or poisoning effects of other materials present, and many other factors. When the fuel and other components are so arranged that a given instantaneous population of neutrons can cause sufficient fissioning to produce at least an equivalent average population of neutrons at later times, the neutron-fission chain reaction is said to be "sustained." The physical facility of which such fuel arrangement is a component part is a nuclear reactor. (In earlier days, the word "pile" instead of "reactor" was frequently used. This originated from the very first nuclear reactor which consisted of a huge "pile" of graphite and uranium.)

Thus, neutrons are the lifeblood of a reactor, just as oxygen is the essential ingredient and controller of chemical combustion of coal in an ordinary furnace. By neutron interactions, the vital process of successive generations of fission in atoms of the fuel, with their attendant release of energy, is carried on. In any given assembly, at any given moment, the rate of fissioning is proportional to the average population of neutrons. When the neutron population is constant, the rate of fissioning is steady, and the power level, i.e., the rate of energy release, is also constant. If conditions are altered so the neutron population and the rate of fissioning increases or decreases, the power level or the rate of energy release of the reactor is correspondingly increased or decreased.

Purposes for which Reactors Are Built. Reactors are usually designed to make maximum use of one or another of the products of fission: heat, radiation, or fission products.

(a) *Utilization of Heat Energy.* Fissioning of

* A neutron source, for example, would be a small capsule containing a mixture of polonium and beryllium. Alphas from the radioactively decaying polonium would interact with the beryllium to yield neutrons: ${}^4\text{He}^4 + {}^9\text{Be}^9 \rightarrow {}^4\text{C}^{12} + n^1$.

1 pound of uranium releases 3.6×10^{10} Btu of heat; or the heat equivalent of 2,000,000 pounds of oil; 1 gram of fissioned uranium yields the heat equivalent of 1 megawatt-day of electricity.

When an atom fissions, most of the heat energy is released instantaneously and essentially at the point of fission. However, most of the energy of the gammas, betas and neutrons which radiate out in all directions is also transformed into heat energy in surrounding materials. Further, the delayed radiation given off by the radioactive fission products is likewise converted into heat energy. Thus, a mixture of fission products continues to generate heat for a long time after the fission events which produced them occurred.

As a source of heat energy for generation of electric power, for desalting of ocean water, for propulsion of mobile vehicles, for the driving force in endothermic chemical processes of industry, and for other similar uses, the heat from nuclear fission has countless potential applications of major dimensions.

(b) *Utilization of Radiation.* Gammas, neutrons, betas and some alphas comprise the radiation given off when an atom is fissioned. The composition of the average radiation is different from one fuel isotope to another. For example, 2.5 neutrons, on the average, are released from each U^{235} fission, while fission of Pu^{239} releases 3.0 neutrons.

The excess neutrons, beyond those needed for sustaining the chain reaction (which is just 1 per fission for maintaining a status quo fissioning rate) and the gamma radiation from fission have many and diverse uses. Some reactors are built for generation and utilization of this radiation (in which case heat may be a nuisance, or an incidental byproduct).

As examples, radiation may be used for:

(1) *Radioisotope production:* Many very useful radioisotopes are produced when excess neutrons from a reactor are captured by ordinary, non-radioactive isotopic species.

(2) *Nuclear fuel production:* Excess neutrons may also be used to convert certain "fertile" isotopic species, which are not useful as nuclear fuels, into readily fissionable nuclear fuels. For example, thorium 232, useless as a nuclear fuel, is transformed by neutron capture and subsequent radioactive decay, into uranium 233. The abundant uranium 238 is transformed by neutron capture from its natural worthless (fuel) value into the extremely valuable plutonium 239.

Thus, by using excess neutrons from nuclear reactors, the potential nuclear fuel supply of the world may be increased many times. U^{235} is the principal naturally occurring nuclear fuel. U^{238} is 140 times more plentiful. Thorium is 7 times as abundant as all the uranium.

(3) *Sterilization of food; chemical catalysts:* The vast gamma and neutron fluxes generated in fission reactors are capable of causing various changes in the physical characteristics of materials, some detrimental and some beneficial, and profound effects in living organisms, including alteration of genetic characteristics.

Foods can be sterilized, and bacteriological processes inhibited; chemical reactions can be catalyzed, and certain types of organic materials can be polymerized by exposure to radiation. The gammas rather than neutrons are principally used in such applications because neutrons may cause undesirable radioactivity in some of the materials.

(c) *Utilization of Fission Products.* Fission products removed from the spent fuel elements of reactors can be used for many purposes. For examples:

(1) Gamma radiation from fission products in bulk has been used for sterilization of food and polymerization of chemicals.

(2) Iodine 132, and various other particular species, have been isolated and used extensively in medical diagnosis and therapy.

(3) Separated, high specific activity isotopes in the multi-megacurie range have been used for radiation therapy sources, for radiography in place of x-rays, and for concentrated sources of heat. For example, Sr^{90} has been used as a concentrated source of heat for direct thermoelectric generation of electricity to be used for instruments in satellites and in remote weather stations.

On the whole, however, to date there are practical uses for only a small fraction of the fission products now being generated, and the chief present concern is for safe and economical storage or disposal of these by-products of fission.

Components of Reactors: Types of Reactors. For a controlled neutron-fission chain reaction in nuclear fuel to proceed, there must be: (1) an assemblage of fuel of appropriate quantity, dimensions and arrangements, (2) a ready means of adjusting the fissioning rate (power level) to any desired level, i.e., a control system; (3) a means of sensing and measuring the processes in progress, i.e., an instrumentation-intelligence system; (4) a means of removing heat from the fuel and disposing of or using it, i.e., a cooling system; (5) a means of protecting the environment from the direct radiation flux of escaping gammas and neutrons, i.e., shielding; (6) a confinement system for retention of the fission products; and (7) other auxiliary components.

In the following sections, some of the basic principles applicable to the functioning of these systems are briefly sketched:

(a) *The Reactor Core.* A reactor core consists of the fuel assembly, of whatever quantity, composition, physical form, dimensions and configuration are required to enable the neutron chain reaction to proceed under the conditions and for the purposes desired in any particular case. In addition, the core includes all other intermixed components, supporting members and immediately adjacent structures and materials.

Reactor fuels are metallic in elemental form but may be readily processed into various compound forms. They may be used in reactors in any of a wide variety of chemical and physical forms. No physical or chemical circumstance affects the fissioning or other nuclear characteristics of any materials.

(b) *Coolant.* When a reactor is to be operated at power levels high enough to generate substantial quantities of heat within the fuel, means must be provided for removal of this heat. Thus, in high-power reactors, fuel is customarily in the solid forms of rods, tubes or plates, having large surface area-to-mass ratios to facilitate escape of heat. Flow channels between adjacent fuel elements accommodate circulation of a fluid heat-removal coolant. In a power reactor, the coolant becomes sufficiently hot to be converted into steam (where water is the coolant) or hot enough to generate steam in the heat exchangers outside the core. The steam then drives turbines to produce electricity in the usual way. Coolants frequently used include air, CO₂, He, N₂, water, organic liquids, sodium, and sodium potassium mixtures.

(c) *Fuel Cladding.* During reactor operation as fissioning proceeds, fission products generated in the fuel tend in many instances to migrate to the surface where, by escape into the circulating coolant, they may be carried to external parts of the system and there give rise to radiation hazards. In most cases, therefore, the fuel itself is clad in a thin impervious, metallic sheath which prevents escape of the fission products. Cladding materials include aluminum, stainless steel, ceramic coatings, and various alloys of aluminum or zirconium.

(d) *Moderator.* When fission occurs, the neutrons released are of very high energy, 2 MeV or so, on the average. It so happens that such high-energy neutrons are inefficient in causing fission in commonly used fuel atoms. Slow neutrons,* however, with energies below 1 eV, have much higher fission efficiencies per collision. Therefore, for most large reactors presently being built, arrangements are made for the fission neutrons to be quickly degraded in energy to slow, thermal energies, and the fission process proceeds mainly by thermal neutron fission. Reactors in which this is the case are called thermal reactors. It is possible and, to achieve certain objectives, it is desirable for the fissioning process to be carried on by the inefficient fast neutron process. Reactors where this applies are called fast reactors.

Neutrons lose energy only, by elastic (i.e., "mechanical," non-nuclear interaction) collision with other atoms. They lose least energy per collision with heavy (e.g., uranium) atoms and most energy per collision with low mass atoms. In thermal reactors, appropriate quantities of low-mass elements such as hydrogen, deuterium, lithium, beryllium, and carbon are intermixed with the fuel to insure that the energetic fission neutrons are promptly degraded in energy, i.e., are moderated in velocity, to the thermal range. Such low mass materials are called moderators. In some cases coolants also serve as moderators.

In fast reactors, all moderating materials are carefully excluded, so the fission neutrons are

forced to collide only with heavy mass atoms and hence lose energy slowly. They eventually do cause fission or may be captured in nonfissionable atoms.

(e) *Reactor Control.* Materials which capture neutrons without subsequently undergoing nuclear fission are called poisons. All materials, even "good" nuclear fuels, are to some extent poisons. Some materials have a much higher non-fission cross section,* i.e., poisoning effect per unit quantity, than do others. Poison materials serve a very useful purpose. They provide, the one most frequently used mechanism, though there are also others, for controlling the fissioning rate or power level of the reactor. In an operating reactor, the judicious insertion or withdrawal of an appropriate quantity of poison can control the rise or fall of the neutron population, and hence the rate of fissioning can be adjusted at will.

In practice, boron or cadmium, in rod form or other suitable shape, attached to drive motors or other suitable remotely manipulated devices, is so arranged that the operator can remove or insert at will just the appropriate amount of poison to cause the neutron population originating from a fixed source to grow rapidly or slowly, remain constant or decrease slowly or rapidly. A reactor is "shut down" when sufficient poison is inserted (or other means are used) to essentially terminate the fission chain reaction.

(f) *Auxiliary Systems.* In addition to the essential components of reactors mentioned above, fuel, coolant, controls and moderator, many other components and systems are necessary for the protection of people and the utilization of the reactor. Appropriate instrumentation is needed to enable the operator to know the status of various processes, neutron flux, radiation levels, temperatures, coolant flow rates, etc. Shielding against radiation fluxes and containment mechanisms to prevent escape of radioactivity are needed for the protection of people. Machinery for collection and utilization of the heat, or to provide access to the radiation fluxes, is necessary to the intended uses of reactors.

A large nuclear reactor facility employed for the generation of electricity or for the propulsion of an oceangoing vessel is exceedingly complicated and expensive.

(g) *Reactors in Use.* Almost 500 reactors have been built, and many more are now under construction in many countries of the world. Some older experimental ones are being dismantled. It is virtually impossible to know the precise numbers in existence at any given time and the various purposes for which they are used. As of August 1964, it was reported (Third International Conference on Peaceful Uses of Atomic Energy) that reactors were being used for production of

* Neutrons in thermal, Brownian equilibrium with surrounding atoms at nominal room temperatures have energies of about 0.25 eV. Such neutrons are called thermal neutrons.

*The "cross-section" of some given nuclear event is the probability per collision that that particular event will occur. For example, an atom of a fuel which possesses a high neutron fission cross section has a high probability of fissioning when a neutron collision occurs.

electricity in various countries approximately as follows: United Kingdom, 2300 MW; United States, 1057 MW; USSR, 900 MW; Italy, 525 MW; France, 150 MW; all others, approximately 70 MW.

As an illustration of the variety and distribution of reactors, Table 1 below lists the reactors built by the United States up to 1964. Table 2 shows the listing of reactors completed as of August 1964.†

TABLE 1. REACTORS BUILT IN THE U.S. 1942* 1964

Type	No.
Experimental, test, research and training	164
Vessel propulsion reactors	52
Production reactors	14
Power reactors	16
Critical facilities	71 cells
<i>Total</i>	317

TABLE 2. REACTORS RECORDED IN COUNTRIES OTHER THAN THE U.S. TO AUGUST 1964

Country	No
Canada	7
France	25
West Germany	13
Italy	17
Japan	13
Netherlands	6
Sweden	6
United Kingdom	58
United States	246
USSR	38
All others recorded	54
<i>Total</i>	483

* The first sustained, controlled nuclear chain reaction was achieved in a graphite natural uranium assembly in Chicago in December, 1942.

CLIFFORD K. BECK

Cross-references: ATOMIC ENERGY, FISSION, FUSION, ISOTOPES, NEUTRON, NUCLEAR REACTIONS, NUCLEONICS, RADIOACTIVITY, REACTOR SHIELDING.

† Foreign reactors built, being built, or planned as of January 30, 1964, Division of International Affairs, USAEC, Washington, D.C.

NUCLEAR STRUCTURE

The basic problem of nuclear structure was first sharply defined in 1932 when Chadwick discovered the neutron. The nucleus was then recognized to be a system of A particles (nucleons) of nearly equal mass (Z protons and $N = A - Z$ neutrons). The forces binding the "nucleons" together must be quite distinct from the well-known electromagnetic interactions which will evidently operate to push the nucleons apart, as well as gravitational forces which appear to be too weak to have any influence in the nucleus at all.

The nuclear forces are found to be of short range ($\sim 10^{-13}$ cm). Within their short radius of interaction, they become strongly attractive (depths near 100 MeV), but at very small distances the interaction becomes so powerfully repulsive that nucleons seldom come closer than about 0.4×10^{-13} cm from one another. This character of the nuclear force produces "saturation." The binding energy per nucleon is nearly constant ($\sim 8.0 \pm 1$ MeV) throughout the periodic table, while each nucleon occupies a volume which may be approximated by a sphere of radius 10^{-13} cm.

The forces described above are assumed to operate between each pair of nucleons. Their exact nature is rather complicated and depends upon the motion, and even the orientation of the nucleons. We do not as yet have a completely trustworthy expression for this force, but rapid progress is being made in this field.

Even if the nuclear force were precisely known, the structure of the nucleus would not be immediately understood in its entirety. Methods are not at present available for solving the three-body problem in closed form for simple interactions, and for more than three bodies, drastic approximations must be used to make the problems tractable. For this reason, considerable attention in physics is devoted to constructing simplified models of the nucleus. The purpose of these models is to isolate the salient characteristics of the nucleus, and thereby obtain some understanding of its behavior.

Two very general classes of nuclear models have received considerable attention. The simplest may be described as "powder models." Here it is assumed that the motion of nucleons within the nucleus is so complicated that statistical mechanics may be employed. In one version of this model, the nucleus is treated in analogy to a drop of liquid (the nuclear forces described above are qualitatively quite similar to forces between the molecules of a liquid). Such a chaotic state seems to ensue if the nucleus is highly excited, so that the nucleons have a great deal of kinetic energy. For this reason, powder models have found their greatest usefulness in describing the nucleus during a nuclear reaction, and especially in describing nuclear fission.

In the shell model (sometimes called the independent particle model), the motion of the individual nucleons is assumed to be much more simple. Each nucleon is, in fact, assigned a definite orbit within the nucleus. The resultant motion is then reminiscent of our usual picture of electrons within the atom, all revolving nearly independently of each other about the nucleus. The motivation for introducing this model into nuclear physics came from the early (1934) observation that nuclei that contained certain numbers of protons or neutrons were unusually stable and abundant. These numbers are generally referred to as magic numbers. The set we recognize today as magic are

N or $Z = 2, 6, 8, 14, 20, 28, 50, 82, \text{ and } 126$

The major details which focus attention on these numbers may be summarized as follows:

(1) *Stability and abundance.* Nuclei with N or Z magic are unusually tightly bound and correspondingly very abundant. If both N and Z are magic (${}^2\text{He}_2$, ${}^6\text{C}_6$, ${}^8\text{O}_8$, ${}^{14}\text{Si}_{14}$, ${}^{20}\text{Ca}_{20}$, ${}^{28}\text{Ni}_{28}$, and ${}^{82}\text{Pb}_{126}$), then the nucleus is even more tightly bound. The last nucleon to complete a magic number has sometimes nearly twice the binding energy of an average nucleon. Furthermore, nuclei with Z magic possess an unusually large number of stable isotopes, and nuclei with N magic have an unusually large number of stable isotones.

(2) *Neutron capture cross sections.* Nuclei with neutron number one short of a magic number have a large neutron capture cross section, while nuclei with a magic number of neutrons exhibit a small neutron capture cross section.

(3) *Islands of isomerism.* Long lived γ -active nuclear states (half-life ~ 1 second) appear just prior to the completion of a magic number.

(4) *Electric quadrupole moments.* Nuclear quadrupole moments tend to be small near magic numbers and large far from magic numbers. A nucleus with N or Z magic appears, therefore, to prefer a spherical shape.

(5) *Delayed neutron emission.* Delayed neutron emitters (e.g., ${}^{86}\text{Kr}_{51}$, ${}^{54}\text{Xe}_{34}$, and ${}^{8}\text{O}_8$) appear when N is one greater than a magic number.

The data clearly point to an interpretation of the magic numbers in terms of shell closures, analogous to the noble gases of atomic structure. Thus it appears that the nucleon-nucleon interaction somehow averages out within the nucleus so that each nucleon, to a reasonable first approximation, sees a fairly smooth potential. Such a potential should be flat near the origin (like a harmonic oscillator), and should go rapidly to zero outside the nucleus (like a square well). A suitable form which has undergone much study is the Fermi function

$$V(r) = V_0[1 + \exp(\alpha(r - a))] \quad (1)$$

In order to reproduce the magic numbers one must add a spin-orbit term:

$$f(r)\mathbf{l} \cdot \mathbf{s} \quad (2)$$

to this central potential. Here l is the orbital angular momentum of a nucleon, and s its spin. The effect of this $\mathbf{l} \cdot \mathbf{s}$ interaction is to couple the orbital and spin angular momenta so that the states with total angular momentum $j = l + \frac{1}{2}$ will be lower in energy than states with $j = l - \frac{1}{2}$. Shell closures obtained with such a potential are found to reproduce the observed magic numbers in a simple and striking manner.

The fact that the shell model works at all in nuclei is somewhat surprising since the theoretical motivation that enhances its success in atomic structure is lacking. One has no strong central field originally in nuclei, and furthermore the nucleons are much more closely packed. Consequently, collisions, which should tend to scatter a nucleon out of its orbit, should be more

frequent. It is found that the Pauli exclusion principle plays a vital role here. When such a collision occurs, the orbit into which a nucleon would tend to be scattered is actually occupied by another nucleon. This fact makes the shell model a far better first approximation than one would originally suspect.

The main effect of the nucleon-nucleon interaction is to produce the average shell model (single-particle) potential. This, of course, is not the sole effect. Many refinements are required before an adequate description of nuclei may be attained.

Consider, for example, a nucleus with just two nucleons beyond a double closed shell. One of the nucleons will go into an orbit j , and the other into an orbit j' . These two orbits may orient themselves such that the nucleus may have total angular momentum J anywhere within the range

$$|j - j'| \leq J \leq j + j' \quad (3)$$

so long as one is consistent with the Pauli exclusion principle. If one only had single-particle potentials all states of this configuration would have the same energy independent of J . To determine the level order actually observed, one must recognize that not all of the original nucleon-nucleon interaction is used up in constructing the single-particle potential. Some "residual interaction" is left over. This residual interaction will still be of the two-body type, and it will remove the degeneracy between states of the same configuration.

The most important feature of the residual interaction is that it produces a pairing force for like orbitals (i.e., $j = j'$). If $j = j'$, and we have like nucleons, the interaction is by far the strongest if $J = 0$. This has the important consequence that for nuclei with both an even number of protons and an even number of neutrons, one must have net angular momentum equal to zero. This rule is never violated. If one adds one nucleon to this even-even nucleus, the result should be a J equal to the j of the odd nucleon. This rule is violated in only a very few cases.

There are a few nuclear phenomena which the shell model cannot describe at all. The sign of the static, nuclear, electric quadrupole moments are generally given correctly by a single-particle model, but the magnitude is frequently found to be larger than that predicted by more than a factor of ten. Similarly, electric quadrupole transition rates are frequently much faster than those given by the shell model. It seems that even in low-lying nuclear levels it is not possible to attribute all of the properties of the nucleus to the nucleons in unfilled orbits. One must take into account possible distortions of the orbits of nucleons in the closed shells.

This line of thought gives rise to the "collective model," in which the nucleons in the core (closed shells) are treated as an incompressible irrotational fluid capable of surface oscillations. Nucleons in unfilled shells will exert a centrifugal force on this fluid and tend to deform it into a

nonspherical shape. In regions of the periodic table where several shell model orbitals are nearly degenerate in energy ($90 \leq N \leq 114$ and $Z > 88$), the nuclear core takes on a rather large spheroidal deformation. The formalism of the theory is reminiscent of that for diatomic molecules. Low-lying excited states are generated through a rotation of the entire system about an axis perpendicular to the axis of symmetry. This gives rise to a band of energy levels:

$$E_J = E_0 + (\hbar^2/2I)J(J+1) \quad (4)$$

where I is a moment of inertia. The identification of spectra in nuclei which could be empirically fitted to Eq. (4) was a great triumph of the collective model.

It should be noted that I is not the moment of inertia of the entire nucleus. It is only the moment of inertia of that part of the nucleus which participates in the rotation, and this is, of course, always less than the "rigid body" value.

We note that the shell model and collective model appear to be quite different in their basic assumptions. The collective model presumes that the nucleon motions are so closely correlated that we can treat a rotation of the nucleus on the whole, while the shell model begins with independent orbitals. It is very important to remember, in this regard, that the shell-model orbitals must be properly coupled together in the nucleus. The Pauli exclusion principle must be satisfied, angular momentum must be a good quantum number, and so on. These orbitals are, therefore not so independent as it might seem. They may very well give rise to collective effects depending upon the exact way in which they are coupled together to obtain the final nuclear wave function.

Various coupling schemes have recently been examined to investigate this, but these are beyond the scope of this article. We present here a simple illustrative example. Consider a set of identical particles attracted by means of identical springs to a common origin. At time $t = 0$ each particle is at the origin and is given some arbitrary velocity. Even though each particle has a different initial velocity, and moves independently of every other particle, the system will undergo a periodic dilation and contraction due to the fact that the frequency of vibration is the same for each particle. Thus we have a collective motion exhibited by a set of independently moving particles.

PAUL GOLDHAMMER

References

- Eisenbud, L., and Wigner, E. P., "Nuclear Structure," Princeton, N.J., Princeton University Press, 1958.
 Elliott, J. P., and Lane, A. M., "Handbuch der Physik, Vol. 39, p. 241, Berlin, Springer-Verlag, 1957.
 Feenberg, E., "Shell Theory of the Nucleus," Princeton, N.J., Princeton University Press, 1955.
 Goldhammer, P., *Rev. Mod. Phys.*, **35**, 40 (1963).
 Inglis, D. R., *Rev. Mod. Phys.*, **25**, 390 (1953).

Mayer, M. G., and Jensen, J. H. D., "Elementary Theory of Nuclear Shell Structure," New York, John Wiley and Sons, 1955.

Cross-references: NUCLEONICS; NEUTRON; PROTON; NUCLEAR REACTIONS; ATOMIC PHYSICS; STRONG INTERACTIONS; WEAK INTERACTIONS.

NUCLEONICS

Nucleonics (nü'kli-on'iks) is a name proposed by Z. Jeffries of the Manhattan District in 1944 to describe the general field of nuclear science and technology. As popularly used, it encompasses not only the study of the atomic nucleus, but also related physical techniques, instrumentation, radiochemistry, and the applications of radioisotopes. In its strict technical sense, it refers only to the first of these and is so used here.

Prior to 1900 atoms were considered to be unalterable and indivisible entities of which the elements were composed. With the discovery of the nuclear transmutations resulting from the radioactive decay process, however, it became apparent that atoms themselves possessed internal structure.

Upon observing that in passing through matter, alpha particles were scattered through larger angles than the then current concepts of the atom would predict, Rutherford suggested that the atom actually consisted of a small, heavy, positively charged nucleus surrounded by negative charges of the same magnitude. Further observations and theoretical refinements have led to the now generally accepted concept that the nucleus also contains elementary particles, i.e., neutrons, which are electrically neutral and have a mass of 1.00897, and protons, which are positively charged and have a mass of 1.00812. Stated in simple form, then, the atom consists of a positively charged nucleus, containing neutrons and protons, surrounded by a number of negatively charged electrons sufficient to provide electrical neutrality. Although this concept of the atom has not been in serious question for almost 50 years, complete understanding of the forces which hold the neutrons and protons together has not yet been achieved.

In 1927, Aston found that experimentally measured isotopic weights differed slightly from whole numbers. From this he was led to the concept of the packing fraction, which is defined as the algebraic difference between the isotopic weight and the mass number, divided by the mass number. Although the theoretical significance of the packing fraction is difficult to assess, it does lead to some interesting conclusions with respect to nuclear stability. A negative packing fraction derives from a situation where the isotopic weight is less than the mass number, inferring that in the formation of the nucleus from its constituent particles, some mass is converted into energy. Since an equivalent amount of energy would be necessary to break up the nucleus into its constituent particles again, a negative packing fraction suggests a high order of nuclear stability. By

the same reasoning, a positive packing fraction indicates nuclear instability. Stable elements with mass numbers above about 175 and below about 25 have positive packing fractions. It is interesting to note that the packing fractions of both hydrogen and uranium are positive.

Actually, a comparison of the isotopic weight with the mass number (as is done in determining the packing fraction) is somewhat artificial. A rigorous determination of the mass-energy inter-conversion in the formation of an atom would seem to require a calculation of the difference between the sum of the masses of the constituent particles of the atom and the experimentally measured isotopic weight. The value of the mass difference thus obtained is the mass defect. The energy equivalent of this mass difference as derived from the Einstein equation yields a measure of the binding energy of the nucleus. Division of the binding energy of a nucleus by the number of nucleons (the total number of protons and neutrons) therein yields the binding energy per nucleon. In stable isotopes, the binding energy per nucleon decreases with increasing mass number, a fact which is important in nuclear fission. Secondly, the binding energy per nucleon derived in the manner described above is an average value, whereas each additional nucleon added to the nucleus has a binding energy less than those which preceded it. Thus, the most recently added nucleons are bound less tightly than those already present.

Additional considerations regarding nuclear stability may be gleaned from a consideration of the odd or even nature of the numbers of protons and neutrons in the nucleus. According to the Pauli exclusion principle, no two extranuclear electrons having an identical set of quantum numbers can occupy the same electron energy state. The application of this principle to the nucleus leads to conclusions which at least are not at variance with observations of nuclear stability. Thus, it is inferred that no two nucleons possessing an identical set of quantum numbers may occupy the same nuclear energy state. It would appear, then, that both protons and neutrons which differ only in their angular momenta or spins may exist in a nuclear state. The exclusion principle requires, therefore, that only protons having opposite spins can exist in the same state. The same consideration applies to neutrons. Accordingly, two protons and two neutrons might occupy the same nuclear energy state provided the nucleons in each pair have opposite spins. Such two-proton-two-neutron groupings are termed "closed shells," and by virtue of their proton-neutron interaction, they confer exceptional stability to nuclei which are made up of them. The nuclear forces in closed shells are said to be "saturated," which means that the nucleons therein interact strongly with each other, but weakly with those in other states. Since like particles tend to complete an energy state by pairing of opposite spins, two neutrons of opposite spin or a single neutron or proton also might exist in a particular energy state.

Any of the above conditions may be achieved when the nucleus contains an even number of both protons and neutrons, or an even number of one and an odd number of the other. Since there is an excess of neutrons over protons for all but the lowest atomic number elements, in the odd-odd situation there is a deficiency of protons necessary to complete the two-proton-two-neutron quartets. It might be expected that these could be provided by the production of protons via beta decay. As a matter of fact, there exist only four stable nuclei of odd-odd composition, whereas there are 108 such nuclei in the even-odd form and 162 in the even-even series. It will be seen that the order of stability, and presumably the binding energy per nucleon, from greatest to smallest, seems to be even-even, even-odd, odd-odd.

Although the existence of binding energies holding the nucleus together has been demonstrated, the problem of defining the nature of these forces presents itself. Clearly, repulsive electrostatic forces must exist between protons. These are "long range" in effect. To achieve nuclear stability then, compensating attractive forces also must exist. It has been concluded that "short-range" attractive forces exist between protons, neutrons, and protons and neutrons. The ($p - n$) attractive forces are considered to be of the greatest magnitude while the ($n - n$) and ($p - p$) forces are of lesser intensity, with the latter decreased by virtue of electrostatic repulsion. When the number of protons in a nucleus is greater than twenty, it is found that the ratio of neutrons to protons exceeds unity. The additional short-range attractive forces provided by the excess neutrons, therefore, may be considered as compensating for the long-range electrostatic repulsive forces between the protons. Nevertheless when the number of protons exceeds about 50, the short-range forces are insufficient to counteract the electrostatic forces completely, with the result that the binding energy per each additional nucleon decreases.

Unfortunately, the nature of the short-range attractive forces between nucleons remains essentially unresolved. An interpretation of them has been presented by Heisenberg, however, in terms of wave-mechanical exchange forces. Thus, if the basic difference between the proton and neutron in a system composed of these two particles is considered to be that the former is electrically charged while the latter is not, then the transfer of the electric charge from the proton to the neutron results in an exchange of individual identity but not a change in the system. That is to say, the system still is composed of a proton and neutron, despite the fact that the particles have exchanged their identities. Since the system itself has the same composition, it must possess the same energy after the exchange as it did before. One of the principles of wave mechanics is that if a system may be represented by two states, each of which has the same energy, then the actual state of the system is a result of the combination, i.e., resonance, of the two separate

states and is more stable than either. In the proton-neutron system under discussion, the energy difference between the "combined" state and the individual states may be considered as the "exchange energy" or "attractive force" between the particles. In an extension of Heisenberg's proposal, Yukawa has postulated that the exchange energy is carried by a new particle which has been given the name meson. Particles having the properties attributed by Yukawa to mesons have been identified in cosmic rays.

With these concepts of nuclear structure and stability, however imperfect, the process of nuclear fission of uranium can now be considered. Although fast neutrons (greater than 0.1 MeV) can cause fission in both uranium 235 and uranium 238, thermal neutrons (about 0.03 eV) are effective only with uranium 235. Uranium 238 is unsatisfactory as a fissionable material for most purposes, however, since it has a high probability for "resonance capture" of fast neutrons, which is a nonfission process. It is instructive to ponder why uranium 235 fissions with thermal neutrons and uranium 238 does not. It will be recalled that the binding energy for an even-even nucleus exceeds that for an even-odd. Consequently, the addition of a neutron to uranium 235, which yields an even-even compound nucleus, will contribute a greater binding energy than in the case of uranium 238 where an even-odd compound nucleus would be produced. Calculations yield a value of 6.81 MeV for the additional neutron in the former case, and 5.31 MeV in the latter. Using Bohr and Wheeler's calculations, it is found that the activation energy for fission is 5.2 MeV for uranium 235 and 5.9 MeV for uranium 238. Thus, the binding energy for an additional neutron in uranium 235 exceeds its fission activation energy, whereas it is less in the case of uranium 238. It can be seen then that uranium-235 fission is energetically feasible with thermal neutrons while the fissioning of uranium 238 is not.

In considering the physical forces acting in fission, use may be made of the Bohr liquid drop model of the nucleus. Here it is assumed that in its normal energy state a nucleus is spherical and has a homogeneously distributed electrical charge. Under the influence of the activation energy furnished by the incident neutron, however, oscillations are set up which tend to deform the nucleus. In the ellipsoid form, the distribution of the protons is such that they are concentrated in the areas of the two foci. The electrostatic forces, of repulsion between the protons at the opposite ends of the ellipse may then further deform the nucleus into a dumbbell shape. From this condition there can be no recovery, and fission results.

It will be recalled that the binding energy per nucleon decreases with increasing mass number. To state it differently, a greater amount of energy is released in the formation of nuclei of intermediate mass number from their constituent nucleons than is the case in nuclei of high mass number. Thus, energy is released in fission because the binding energy of the high mass number

uranium-236 compound nucleus is less than that of the intermediate mass number fission products which are produced. The total energy thus liberated in fission is about 200 MeV. Of this the kinetic energy of the fission products accounts for 160 MeV. These fragments, being of significantly lower atomic number, require fewer neutrons for stability than they actually contain immediately after fission. These excess neutrons, therefore, are "boiled off" the fission fragments, the process occurring in two distinct phases. In the first phase, "prompt" neutrons of about 2-MeV energy are released within 10^{-12} seconds after fission occurs and take up about 7 per cent of the fission energy. Subsequently, after several seconds, additional "delayed" neutrons with about 0.5-MeV energy are boiled off the fission products. These play an extremely important role in the control of the chain reaction in nuclear reactors.

Measurements made by Zinn and Szilard indicate that an average of 2.3 neutrons are released per fission. If in each fission, therefore, at least one of the released neutrons caused the fissioning of another uranium nucleus, there would result a series of fissions, i.e., a self-sustaining chain reaction. The establishment of this condition basically is dependent upon two related factors: (1) the number of neutrons which are lost by escape from the geometrical confines of the system and (2) the number which are used up in nonfission processes, such as capture by uranium 238. Loss by escape is proportional to the surface area of the system. A sphere, for example, would provide a maximum volume of fissionable material for a minimum surface area. Loss by capture may be controlled by the use of uranium suitably enriched in the 235 isotope. Thus, the quantity, quality and geometry of the system may be varied until a "critical" mass is obtained. Under these conditions, there is a constant number of neutrons in the system, that is to say, the "reproduction factor" is unity.

In order that the chain reaction might proceed under controlled conditions, certain additional features must be built into the system. Thus, the neutrons must be slowed down so that the thermal neutron uranium-235 fission process will be maximized and fast neutron uranium-238 nonfission capture will be minimized. This may be accomplished by causing the fast neutrons to undergo a series of collisions, say with carbon atoms, until their energy has been reduced below the excitation energy for the fast fission process and the threshold energy for nonfission capture. In addition, some means must be had to control the reproduction factor. The introduction of variable amounts of a strongly neutron-absorbing material, such as boron, will serve this purpose. Lastly, protection must be afforded against the lethal radiations accompanying the chain reaction.

All of the above features are incorporated in the nuclear reactor. In the typical graphite reactor, a "pile" of graphite blocks are stacked atop each other in the form of a cube. Through

each block, a hole is bored, into which the uranium fuel rods are placed. Interlaced between the graphite blocks are boron control rods whose depth of penetration into the pile may be varied. Around the whole structure a several foot thick concrete shield is poured. In operation then, the fast neutrons produced by fissioning of the uranium are thermalized in the graphite blocks. The number of thermal neutrons in the system, and hence the reproduction factor, is controlled by varying the depth of the strongly neutron-absorbing boron control rods. The radiations are absorbed in the concrete shield.

It will be recalled that the prompt neutrons are released within 10^{-12} second. Manipulation of the control rods to prevent an excessive buildup in the power level would be difficult in view of the small latitude permitted by this time interval. The delayed neutrons are emitted only after several seconds, however, and thus provide the required time for the operation of the control rod mechanism.

The graphite reactor is an example of the "heterogeneous" type reactor. In this system the fuel and the moderator (graphite) are fabricated separately in rods and blocks and stacked in a lattice form. With "homogeneous" reactors, on the other hand, the fuel and moderator are intimately mixed, as in a solution or slurry.

In the operation of a nuclear reactor, both heat and radioactive isotopes are produced. The latter may be obtained as byproducts from the processing of fuel elements to remove the fission products or by insertion of a target material in the reactor for neutron irradiation. Where the radioactive product is isotopic with the target, it usually is not possible or practical (and for most purposes not necessary) to separate the two species. If the product is not isotopic with the target, of course, chemical separation is feasible. Radioactive materials find wide use in basic and applied research, industrial applications, agricultural studies, and medical research, diagnosis, and therapy.

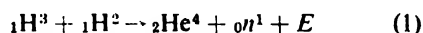
The large amount of heat produced in the nuclear reactor immediately suggests the possibility of using the reactor for the production of electric power. The basic change over conventional power stations involves only the replacement of the coal furnace by the reactor as a source of heat. The turbines, transmission lines, etc., would remain unaffected. Although a nuclear power station is feasible in principle, a number of technical and economic considerations present themselves. The former involve the development of suitable structural materials for the reactor and efficient heat transfer methods. In addition, the design of the reactor power station must incorporate features which will minimize the danger to surrounding communities in the event of an accident. Provision also must be made for safe disposal of the radioactive waste products (see NUCLEAR REACTORS).

In 1932, Cockcroft and Walton, using an accelerator, bombarded lithium 7 with protons and obtained helium 4. The reaction was accompanied by an energy release of 17.3 MeV

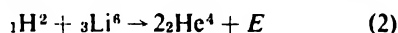


This was the first demonstration of nuclear fusion, i.e., the production of elements of higher atomic number (e.g., He) from elements of lower atomic number (e.g., H). As in the case of the fission reaction, energy is released because the binding energy of the proton is less than that of the alpha particle produced.

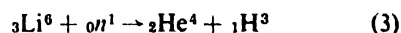
Unlike the fission reaction, in which there is no electrostatic repulsion to the approach of the uncharged neutron to the target nucleus, the fusion reaction requires the coalescence of two positively charged nuclei. It is clear that at least one of the reacting particles must possess exceedingly high energy to overcome this potential barrier. A temperature of about 20 million degrees centigrade is necessary to effect the reaction:



while even higher energy is required in the case of



Although the Atomic Energy Commission has not released information concerning "thermonuclear devices," discussions of them available in the open literature are of interest. Several writers have postulated that a conventional fission bomb provides the energy necessary to initiate reaction (1) and this in turn furnishes sufficient energy to trigger reaction (2). These authors believe it would not be feasible to employ a sufficiently large initial charge of tritium to make a practical device and therefore conclude that the neutron produced in reaction (1) reacts with the lithium 6 employed in reaction (2) to yield further quantities of tritium:



Assuming this to be the case, only a small initial charge of tritium would be necessary since the additional quantities produced in reaction (3) would be available to react with deuterium as per reaction (1). Thus, they envisage a type of "chain reaction" involving the sequence $1 \rightarrow 2 + 3 \rightarrow 1$. Work is in progress in several countries on a controlled thermonuclear reaction.

JOSEPH E. MACHUREK

Cross-references: ATOMIC PHYSICS, NUCLEAR RADIATION, NUCLEAR REACTIONS, NUCLEAR REACTORS, NUCLEAR STRUCTURE, RADIOACTIVITY, STRONG INTERACTIONS, WEAK INTERACTIONS.

O

OHM'S LAW. See CONDUCTIVITY, ELECTRICAL.

OPTICAL INSTRUMENTS

Optical instruments are of two main types: those which project a real image, e.g., the projection lantern and the camera, and those in which the eye views a virtual image, e.g., the microscope and telescope.

The main features of optical instruments can be conveniently considered under the concepts which have been devised in their study.

Magnifying Power or Angular Magnification. This is the ratio of the angle subtended at the eye by the image to the angle subtended at the eye by the object, when the latter is in the most favorable position for observation.

Microscopes. For a normal eye, an object cannot be brought nearer than 25 cm for distinct and comfortable vision. A single converging lens, of focal length f cm, forming a virtual image at the least distance of distinct vision, has a magnifying power of $25/f + 1$. Thus a lens of focal length 2.5 cm. has a magnifying power of 11 and this is about the maximum conveniently attainable with a single lens since aberrations increase considerably as the focal length becomes shorter and a short-focus lens has to be held very close to the object (see LENS).

In the (compound) microscope a converging lens, called the objective, produces a real inverted image with a magnification of v/u , where v and u are the distances of the image and object respectively from the lens (Fig. 1). An eyepiece, acting like a magnifying glass, further magnifies this image, giving a combined magnifying power

of $v/u (25/f_2 + 1)$. The magnification produced by the objective could be made large by making v large and u small. However v is limited by the length of the microscope tube—usually 16 cm. The maximum useful magnifying power of a microscope is about 2000 because of the limited resolving power of the instrument. There is little point in increasing the size of the final image if no further detail in the object is thereby revealed (see MICROSCOPE).

Telescopes. In the case of the telescope, the object cannot normally be brought closer to the eye, and the magnifying power is the ratio of the angle subtended at the eye by the image to the angle subtended at the eye by the object. When the telescope is in normal adjustment (with the final image at infinity), the magnifying power is f_1/f_2 , where f_1 and f_2 are the focal lengths of the objective and eyepiece respectively.

If the objective of a simple astronomical telescope, consisting of two converging lenses, is illuminated with a ground glass screen and a lamp, the eye lens will form an image of the objective which can be focused on a sheet of paper. This image is known as the *exit pupil* (Fig. 2), and all rays leaving the instrument must pass through it. A metal cap with a circular hole is usually placed here in an actual instrument. The *entrance pupil* determines the amount of light which can enter the instrument and in this case it is the objective. When the telescope is in normal adjustment,

$$\text{Magnifying power} = \frac{\text{Diameter of entrance pupil}}{\text{Diameter of exit pupil}}$$

Astronomical telescopes are classified by the diameters of their objectives because this deter-

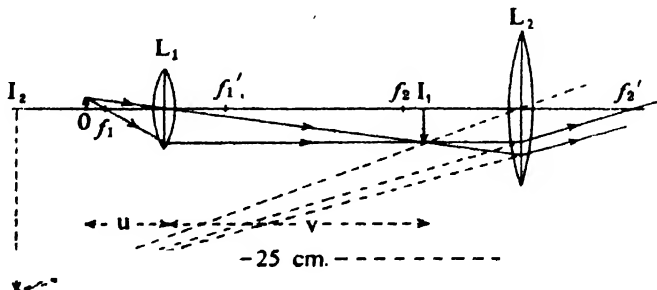


FIG. 1

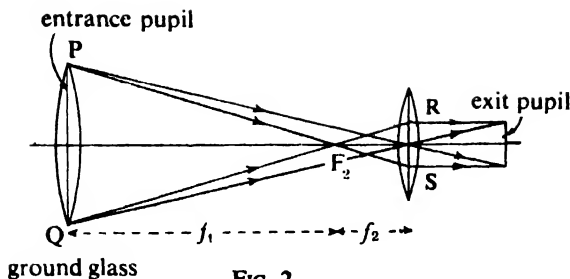


FIG. 2

mines their resolving power. At night, the pupil of the eye has a diameter of about 8 mm or $\frac{1}{3}$ inch, and hence the diameter of the exit pupil of the telescope should be the same. In the case of a 6-inch telescope.

$$\text{Magnifying power} = \frac{6}{\frac{1}{3}} = 18$$

A lower power than this should not be used because the effective aperture of the objective would be reduced.

The ratio of the diameter of the objective to its focal length, known as the *relative aperture*, determines the amount of aberration in the image. If it is decided to use an $f/15$ objective, its focal length must be $6 \times 15 = 90$ inches. Hence the focal length of the eyepiece must be $90/18 = 5$ inches. Thus the focal lengths of the lenses in a telescope are controlled by the diameters of the objective and of the pupil of the eye and also by the relative aperture of the objective.

Resolving Power and Limit of Resolution. The resolving power of an instrument is its ability to reveal fine detail.

Microscopes. Abbe showed that the least distance, x , between two object points which can just be resolved by the objective of a microscope, called the limit of resolution, is given by

$$x = \frac{\lambda}{2\mu \sin \alpha}$$

where λ is the wavelength of the light used, μ is the refractive index of the medium between the object and objective, and 2α is the angle subtended by the diameter of the objective at the object points (Fig. 3); the term $\mu \sin \alpha$ is called the *numerical aperture*. When the object is in air and $\mu = 1$, since the maximum value of $\sin \alpha$ is 1, the maximum limit of resolution of the objective is $\lambda/2$. Thus two object points would just be resolvable if separated by at least half the wavelength of the light used.

For the highest resolving power the angular aperture, 2α , of the objective of a microscope must be as large as possible. Hence the object must be brought as close as possible to the front surface of the objective. This is achieved by constructing objectives of very short focal length ($2/3$, $1/6$ or $1/12$ inch).

The resolving power can also be increased by increasing the value of μ in the above formula;

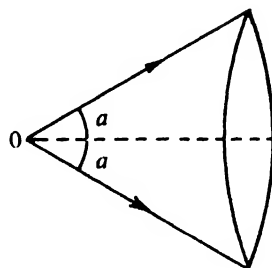


FIG. 3

cedar oil is placed between the cover slip and the objective, an arrangement known as the *oil immersion objective* (Fig. 4).

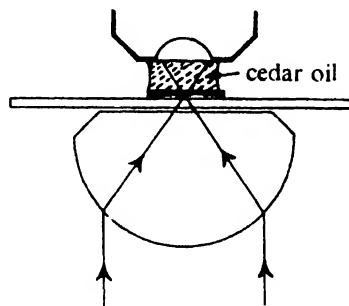


FIG. 4

Yet another way of increasing the resolving power is to reduce the value of λ , and this is done in the *ultraviolet microscope* employing radiation of wavelength about half that of visible light. The resolving power is thereby doubled and the useful magnification is likewise doubled. This microscope must be combined with a camera to make the image visible.

The *reflecting microscope*, still in process of development, in which the objective is a concave mirror, also is used mainly with ultraviolet light. One of the chief advantages of the instrument is its long working distance, i.e., the distance between the object and the objective.

Electrons may be focused by electric or magnetic fields and hence may be used in a microscope instead of light. Electrons behave like waves and the wavelength is shorter the greater their speed. When accelerated under a potential

difference of 60 kV—one commonly used in *electron microscopes*—the electrons have a wavelength only 1/100,000th that of light, giving a resolving power 100,000 times greater (see ELECTRON MICROSCOPE).

Telescopes. The angle, ϕ , subtended at the objective of a telescope by two points on an object which can just be resolved, called the limit of resolution, is given by

$$\phi = \frac{1.22\lambda}{a}$$

where λ is the wavelength of the light and a is the diameter of the objective. Hence, to obtain a high resolving power, the diameter of the objective must be made as large as possible. Furthermore, the full aperture of the objective must be utilized by ensuring that the exit pupil is not greater than that of the pupil of the eye.

The Yerkes telescope, the largest refractor in the world, has an objective of diameter 40 inches. Taking $\lambda = 6 \times 10^{-5}$ cm as the average wavelength of white light,

$$\phi = \frac{1.22 \times 6 \times 10^{-5}}{40 \times 2.54} \quad (1 \text{ inch} = 2.54 \text{ cm})$$

$$= 0.15 \text{ second of arc}$$

The same expression for the limit of resolution applies to concave mirrors as well as to converging lenses. The largest concave mirror objective, at Mount Palomar, has a diameter of 200 inches, five times that of the Yerkes objective, and hence its limit of resolution is $0.15/5 = 0.03$ second of arc.

By comparison, the limit of resolution of the unaided eye is about 1 minute of arc.

Depth of Focus and Depth of Field. The greatest distance the plate of a camera can be moved without spoiling the definition of the image is called the depth of focus. The corresponding distance between the positions of the object at which it is sufficiently in focus on the plate is called the depth of the field (also called depth of focus).

If a point object at a particular distance from a camera lens is exactly focused, so that its image on the plate is a point, then another point object at a different distance from the lens will give rise to an image consisting of a blurred circular patch. Since the visual acuity of the eye

is about 1 minute of arc, the distance apart, x cm, of two points which are just distinguishable at the least distance of distinct vision, 25 cm, is given by

$$\frac{x}{25} = 1 \times \frac{\pi}{180 \times 60} \text{ radians}$$

$$x = 0.007 \text{ cm.}$$

Two circular patches, of diameters 0.007 cm, with their centers 0.007 cm apart, would, of course, just touch each other.

For contact prints, a circle of diameter 0.025 cm is taken instead of 0.007 cm, as a sufficiently small image of a point object, but this is clearly too big if the photograph is to be enlarged.

The depth of focus of a lens can be increased by reducing its aperture with a stop. In Fig. 5, the diameter of the patch of light on the plate is reduced by the stop and the image of O will be reasonably clearly focused if the diameter does not exceed 0.025 cm.

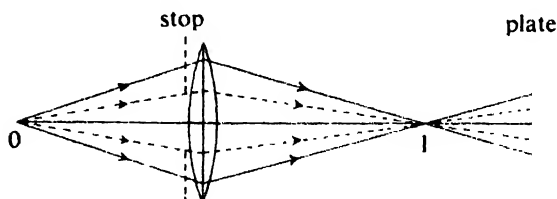


FIG. 5

The depth of focus of a microscope is approximately inversely proportional to the square of the numerical aperture. An objective of numerical aperture 1.40 has, at a magnification of 1000X, a depth of focus of only 0.0005 mm.

Field of View. The field of view of an instrument is the angle subtended at the eye by the largest object, the whole of which can just be seen. This angle, in the case of a simple telescope consisting of two converging lenses, is $a_2/(f_1 + f_2)$, (Fig. 6), where a_2 is the aperture of the eye lens and f_1 and f_2 are the focal lengths of the objective and eye lens respectively. If f_1 is doubled, thereby doubling the magnifying power, the field of view is nearly halved (f_2 is small compared with f_1).

The falling off in brightness of the edge of the image, owing to some rays from the objective failing to strike the eye lens, is called *vignetting*,

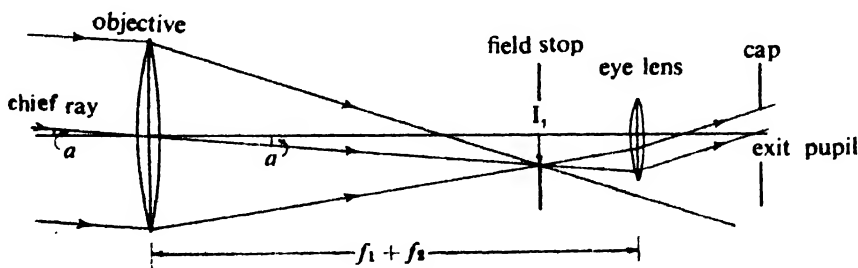


FIG. 6

and it is avoided or kept to a reasonable value by placing a circular opening, called the *field stop*, of suitable size where the image formed by the objective is situated.

Illumination in Microscopes. The final image in a microscope magnifying 1000X has an area one million times that of the object. It is therefore only one-millionth as bright, assuming no light loss in the instrument. Light is concentrated on the (transparent) object by means of a *substage condenser*.

Dark ground illumination is used for viewing tiny particles or very fine lines. The illumination is made too oblique for light to pass direct from the condenser into the objective. The objects are seen by scattered and diffracted light against a dark background.

Phase contrast illumination is a method of making the structure of transparent objects visible. The arrangement consists of an annular stop at the condenser and, beyond the objective, a phase plate consisting of a glass plate on which there is an annular layer of material to increase the optical path of the light by a quarter of a wavelength. Interference between direct and diffracted light augments the diffraction pattern of the image.

Interference microscopy is another method of making fine structure visible. The transparent object is placed between two semi-silvered surfaces and the interference between light passing through the object and light passing by it is observed.

Polarized light, obtained by a suitable polarizer below the condenser, is used in microscopes designed for geological testing of constituents in rock specimens.

A. E. E. MCKENZIE

References

- Cooper, H. J., Ed., "Scientific Instruments," Vols. I and II, London, 1946 and 1948.
Martin, L. C., "Technical Optics," Second edition, Vols. I and II, London, 1961.

Cross-reference: ABERRATIONS; LENS; MICROSCOPE; OPTICS, GEOMETRICAL; REFLECTION; REFRACTION.

OPTICAL PUMPING

The term "pompage optique" was coined by A. Kastler to describe the process of "pumping" atoms from one hyperfine quantum state to another by a process of resonant fluorescent scattering of light. The original purpose was to facilitate the detection and measurement of the radio-frequency (rf) fine and hyperfine structure, which otherwise is observable by magnetic resonance and atomic beam methods (see MAGNETIC RESONANCE, also see ATOMIC AND MOLECULAR BEAMS), yet which cannot be directly observed by optical spectroscopy, primarily because of the Doppler width of spectral lines. Optical pumping has now come to embrace any experimental work

in which fine structure, rf spectroscopic measurements, polarization of electrons or nuclei, atomic cross sections, and oscillator strengths are measured or produced by means of polarized, filtered, or modulated light, and perhaps detected by the light as well. The distinction between measurements made in this way and those of conventional spectroscopy arises from the fact that the wavelength of the light is not used to measure energy splittings.

A sample description of the process may be given in schematic form. Consider a sample of atoms in gas or vapor phase, which have a $^2S_{1/2}$ ground state. Sodium is a good example, but we ignore the nuclear spin. If exposed to the resonance radiation coupling this ground state to the $P_{3/2}$ and $P_{1/2}$ states, the atom will absorb and reemit rapidly, and if the absorbed radiation is circularly polarized, the atoms will absorb one unit ($\hbar/2\pi$) of angular momentum with respect to a fixed axis of quantization (or lose one depending on the sign of the circular polarization) as shown in Fig. 1. If the atom is not disturbed during the

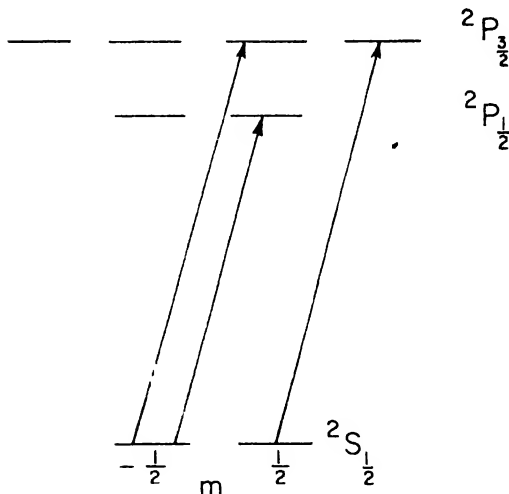


FIG. 1. Absorption of circular polarized resonance radiation by a typical alkali metal.

time it is in the excited state, it will reemit any polarization, but on the average will return to the ground state with one higher (or lower) unit of angular momentum in the quantization direction. This polarization is that of the electron, but if the nucleus has a spin it is so tightly coupled to the electron spin at low fields and in the radiation process that the resulting polarization is that of the combined nucleus and electron according to the angular momentum coupling rules for $\vec{I} + \vec{J} = \vec{F}$. The polarization may be measured by disorienting the atoms by adiabatic field direction changes, or by inducing transitions between the various hyperfine or Zeeman levels and measuring the transmitted or scattered light.

If the sample vapor is mixed with a "buffer gas," which helps to prevent the diffusion of the

sample vapor out of the light beam and reduces the Doppler width of the hyperfine lines, then the atoms may be disoriented in the excited state and the pumping is reduced. However, if filtering is used, it is possible to maintain orientation effects even at several atmospheres of the buffer gas. In the example mentioned, removal of the light which couples the $S_{\frac{1}{2}}$ to the $P_{\frac{1}{2}}$ state means that the absorption probability of one of the $S_{\frac{1}{2}}$ states is reduced to zero and so equilibrium is established with the least-absorbing state most populated.

Another example will show the use of the technique in excited states. A sample such as mercury (or other Group II elements) is exposed to its resonance radiation which is linearly polarized parallel to axis of quantization. This light will cause transitions from the 1S_0 ground state to the 3P_1 $m = 0$ state. The scattered light will have the same polarization. However, if magnetic dipole transitions between the excited state levels of different m , as shown in Fig. 2, are induced by

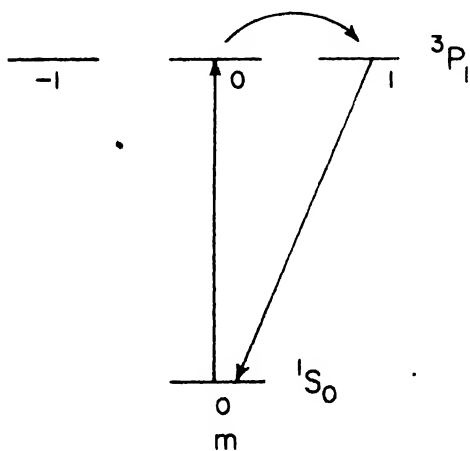


FIG. 2. Absorption and scattering of plane polarized resonance radiation by mercury vapor when excited-state transitions are made.

rf fields, circular polarization will also be emitted and may be readily detected by a polarization analyzer. If the mercury nucleus has a spin, the ground state will also be polarized in the manner described in the first example, and the measurements are equivalent to observation of nuclear magnetic resonance in a gas, but detected by light.

A typical experimental arrangement to observe optical pumping is shown in Fig. 3. A powerful and non-noisy resonance radiation lamp is focused through polarizers and filters onto the sample cell, which may be heated to get the required vapor pressure of sample. An axis of quantization is established by a magnet, solenoid or Helmholtz coils. Transmitted or scattered light is collected and monitored by a photodetector. To facilitate detection it is often desirable to make

use of modulation techniques, similar to those used in magnetic resonance experiments (see MODULATION and MAGNETIC RESONANCE), in which the magnetic field or radio frequency is modulated, and the resulting alternating current signal synchronously detected or amplified by a phase sensitive amplifier. Because of the high efficiency in the scattering and detection process with light quanta, optical pumping signals are usually much stronger than could be obtained by direct radio-frequency or microwave absorption experiments on comparable samples.

The optical pumping of many isotopes of the Group I and II elements has been studied. It is difficult to polarize atoms in gas cells which do not have an S ground state because collisions will rapidly disorient the orbital angular momentum of the electrons, but these atoms may be studied in a beam. The process cannot be used in most cases at sample pressures higher than 10^{-3} or 10^{-4} torr, because the light will not penetrate such dense samples.

It is not possible to describe briefly a general theory of optical pumping. Each system has its own spectroscopic splitting which must be analyzed. It is usually necessary to write a "master equation" to determine the equilibrium populations in the sublevels, taking into consideration the transition probabilities into and out of these levels, and the relaxation rates. Fortunately, in most cases, all that is needed is knowledge of the relative transition probabilities, and these will probably depend only on the angular momentum quantum numbers.

One of the most effective means of extending optical orientation to other elements is by means of the spin exchange process. Outer electrons will overlap during the collision of atoms, and the spins will interchange as a consequence of a "beat" between symmetric and antisymmetric spin configurations. In the case of S -state wave functions, the coupled angular momentum is conserved, and the polarization produced by optical pumping on one atom will be thermodynamically shared with all other atoms with which a spin exchange is possible. Since outer electrons may overlap at interatomic distances much greater than a collision impact parameter, the spin exchange cross section is usually much greater than the collision cross section. Atomic hydrogen and atomic nitrogen, as well as the alkali metals, have been studied by these methods. In fact, optical orientation and spin exchange may be expected to work on ions, and the first observation of spin exchange was made with the free electron. Polarized electrons were removed by ultraviolet or discharge ionization from an optically pumped sodium vapor sample. The electrons were disoriented by rf fields, and subsequently disoriented the sodium by exchange in an observable way.

Other methods of coupling optically produced orientation also exist. Metastable (3S_1) helium atoms are long lived and may be optically pumped. The metastability is transferred with a reasonably large cross section to ground state (1S_0) atoms, and if the helium is the isotope He^3 ,

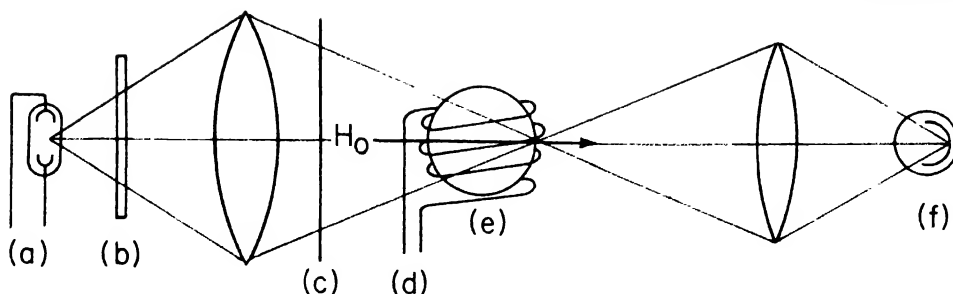


FIG. 3. Typical optical pumping apparatus: (a) resonance radiation lamp, (b) filter, (c) polarizer, (d) rf coils, (e) sample cell, (f) photodetector

the nuclear polarization is also efficiently transferred to the ground state, making it possible to produce at pressures of several torr, samples of 40 per cent polarized He^3 with relaxation times of about one hour. The contact hyperfine interaction and dipolar interactions also can transfer the angular momentum of optically oriented atoms to other atoms, but cross sections are much smaller.

Several interesting light modulation effects may be observed in pumping experiments. Light beat and modulation effects occur when an atom scatters light while it is simultaneously undergoing hyperfine transitions. If a sample is oriented in the Z direction, and a rf field excites a Zeeman resonance, the atoms will precess coherently. An additional polarized light beam passing through the sample will be modulated at the Larmor frequency. As an inverse effect, if the pumping light itself is modulated at the appropriate radio frequencies, transitions will be induced. In the example of mercury, if magnetic dipole transitions are induced in the sublevels of the (3P_1) states, the scattered light will show the beat frequency modulation of the sublevels. The varieties of these modulation effects are by no means yet exhausted.

A useful effect for studying hyperfine structure in cases where the system cannot easily be pumped is the "level-crossing" effect. Some of the levels of the fine and hyperfine structure of states of atoms more complicated than the examples described here cross one another at particular values of a perturbing magnetic field. At these points, the optical opacity of the sample to the appropriate light will be changed, and from the values of the magnetic field, and a knowledge of the general form of the hyperfine splitting, magnetic moments and hyperfine energies may be determined.

There are numerous practical applications of these methods. Optically pumped gas cell magnetometers have been made utilizing rubidium vapor or metastable helium (see MAGNETOMETRY). If modulated light is fed back to rf coils, a self-contained oscillator may be built. Rubidium and cesium gas cell ATOMIC CLOCKS have been built with stabilities of 1 part in 10^{11} . These may be made in a very simple manner, utilizing those hyperfine transitions which are not

perturbed by magnetic fields. Atomic gyroscopes may be constructed by using the optically pumped angular momentum and the light to detect precession. Perhaps the most important application has been to produce population inversion required in the operation of masers and lasers, particularly potassium and rubidium vapor, and chromium in ruby (see LASER and MASER).

THOMAS R. CARVER

References to Review Articles

- Bloom, A. L., *Sci. Am.*, **203**, 72 (1960).
 Carver, T. R., *Science*, **141**, 599 (1963).
 De Zafra, R., *Am. J. Phys.*, **28**, 646 (1960).
 Kastler, A., *J. Opt. Soc. Am.*, **47**, 460 (1957).
 Series, G. W., *Rept. Progr. Phys.*, **23**, 280 (1959).
 Skrotskii, G. V., and Izyumova, T. G., *Soviet Phys.-Usp. (English transl.)*, **4**, 177 (1961).

Cross-references: ATOMIC AND MOLECULAR BEAMS, ATOMIC CLOCKS, ELECTRON SPIN, LASER, MAGNETIC RESONANCE, MAGNETOMETRY, MASER, MODULATION, POLARIZED LIGHT, SPECTROSCOPY, ZELMAN AND STARK EFFECTS.

OPTICS, GEOMETRICAL

The radiant energy emitted from a point of a luminous source situated in a homogeneous medium free from obstacles travels through the medium as a spherical wave front whose velocity of advance is a characteristic of the medium. A radius of this wave front is a light ray. If a ray experiences a change of medium or encounters an obstacle, it will, in general, deviate from the rectilinear path defined by the radius. The science of geometrical optics is concerned with controlled deviations.

The Law of Reflection. (a) The directions of incidence and reflection and of the normal to the reflecting surface are coplanar.

(b) The angle of reflection is equal to the angle of incidence, both angles being taken with respect to the normal (see Fig. 1(a)).

The Law of Refraction. The directions of incidence and refraction and of the normal to the refracting surface are coplanar.

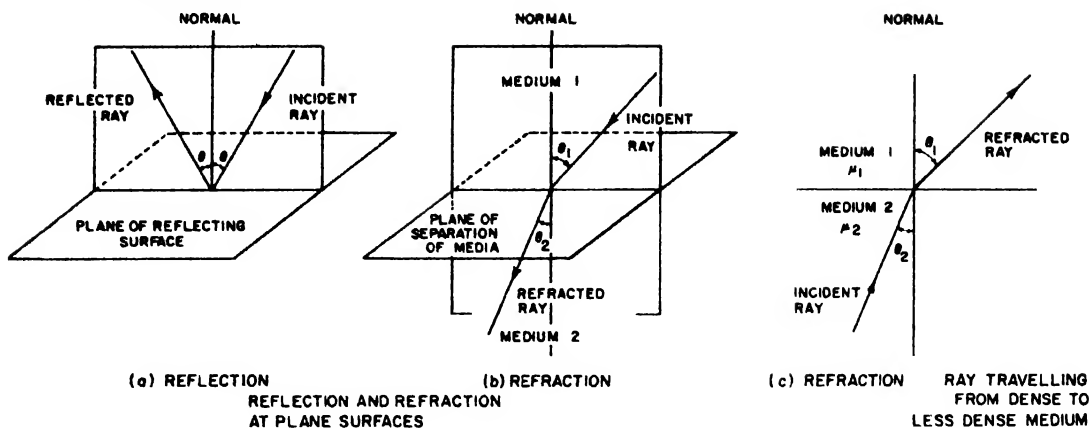


FIG. 1. a-c Reflection and refraction at plane surfaces.

(b) The ratio of the sines of the angles of incidence and refraction, both angles being taken with respect to the normal, is a constant whose magnitude is a function of the properties of the two media on either side of the refracting surface (see Fig. 1(b)). The function is

$$\sin \theta_1 / \sin \theta_2 = \mu_{12}$$

The constant μ_{12} is the index of refraction for the two media when light travels from medium 1 to medium 2. If the direction of the ray is reversed, as in Fig. 1(c), it traverses exactly the same path so that

$$\sin \theta_2 / \sin \theta_1 = \mu_{21}, \quad \therefore \mu_{21} = 1/\mu_{12}$$

If the first medium is air, it is usual to drop the suffixes and to replace μ_{12} by n . If the first medium is a vacuum then μ_{12} is the absolute refractive index of the second. For air at NTP the value is 1.00028 so that, for most practical purposes the distinction between μ_{12} and n , with air as the first medium, is negligible for substances, such as glass, commonly used for light control. The value of n depends on the wavelength of the light. For

sodium yellow of 5890.6\AA , the values of n for a number of media are:

Crown glass	1.517
Flint glass	1.650
Quartz	1.544
Fused silica	1.459
Water at 20°C	1.333
Diamond	2.42

For a mathematically plane surface, the coefficient of reflection for normal incidence is dependent on the value of μ_{12} . Fresnel's law gives for this coefficient

$$\rho = [(\mu_{12} - 1)/(\mu_{12} + 1)]^2$$

Clean polished surfaces give results largely in accord with Fresnel's law. Films of grease or slight roughness appreciably reduces the value of ρ .

Total Reflection. Figure 2(a) shows a series of rays passing from air into a denser medium. Ray No. 1 experiences no deviation. The others experience progressively increasing deviations until, with the tangential ray, No. 4, the angle of

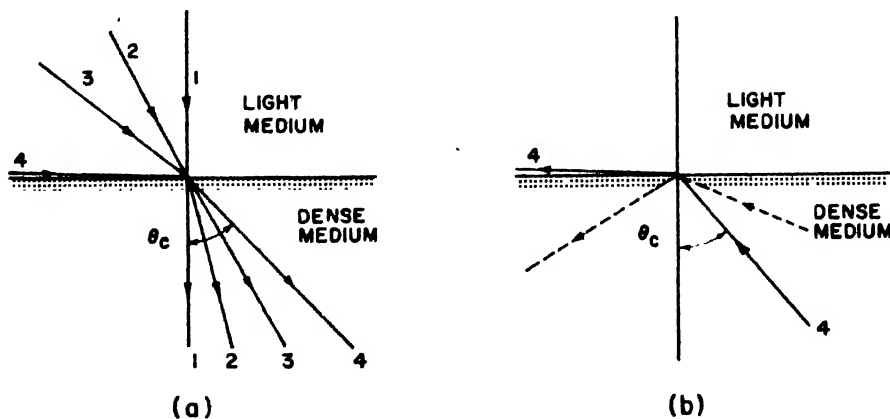


FIG. 2. Illustrating total reflections.

refraction θ_c is the greatest possible for that particular medium. Since the angle of incidence is 90°

$$1/\sin \theta_c = n, \quad \therefore \sin \theta_c = 1/n$$

By reversing the directions of the rays, we see that an angle of incidence θ_c is the maximum possible angle for emergence into the air. For angles $> \theta_c$ there is internal reflection, as in Fig. 2(b), governed by the law of reflection. The angle θ_c is called the critical angle for the particular combination of media. Figure 3 shows a number of practical applications of total reflection.

Image Formation. The geometry of Fig. 4(a) shows that the image in a plane mirror of a point in front of it is another point which lies behind the mirror along the normal from the object point, and is as far behind the mirror as the object is in front. The image of an extended object is identical in size and shape but experiences lateral inversion (see Fig. 4(b)). The image is virtual in the sense that it cannot be received on a screen.

The geometry of Fig. 4(c) shows that the image of a point in a dense medium is another point situated on the same normal to the surface and at a distance from the surface of $1/n$ of the distance of the object point. For large viewing angles, this relationship does not apply, the rays forming the envelope of a caustic as in Fig. 5(a). If the object is extended, a straight line for example, the

image is curved and concave towards the surface as in Fig. 5(b).

Refraction by Prism. Figure 6(a) shows the general case, the incident angle θ_1 being different from the emergent angle θ_1' . The deviation D is given by

$$D = \theta_1 + \theta_1' - A$$

Figure 6(b) shows the conditions for minimum deviation. $\theta_1 = \theta_2 = \theta$, say. The deviation is now given by

$$n = \sin \frac{A + D_{\min}}{2} / \sin \frac{A}{2}$$

Thus for a 60° prism for which $n = 1.517$

$$\sin \frac{A + D_{\min}}{2} = n \sin A/2 = 1.517 \times 0.5 = 0.7585$$

$$\therefore A + D_{\min} = 2 \times 49^\circ 20'; \quad D_{\min} = 38^\circ 40'$$

A "thin" prism is one for which $A \gg 15^\circ$, so that θ (radians) $\approx \sin \theta$. This gives

$$D_{\min} \approx (n - 1)A$$

Reflection at Curved Surfaces. The element of area round a point on the surface is regarded as part of the tangent plane at that point. First consider a spherical surface, Figs. (7a) and (b), and let the aperture be small compared with the radius

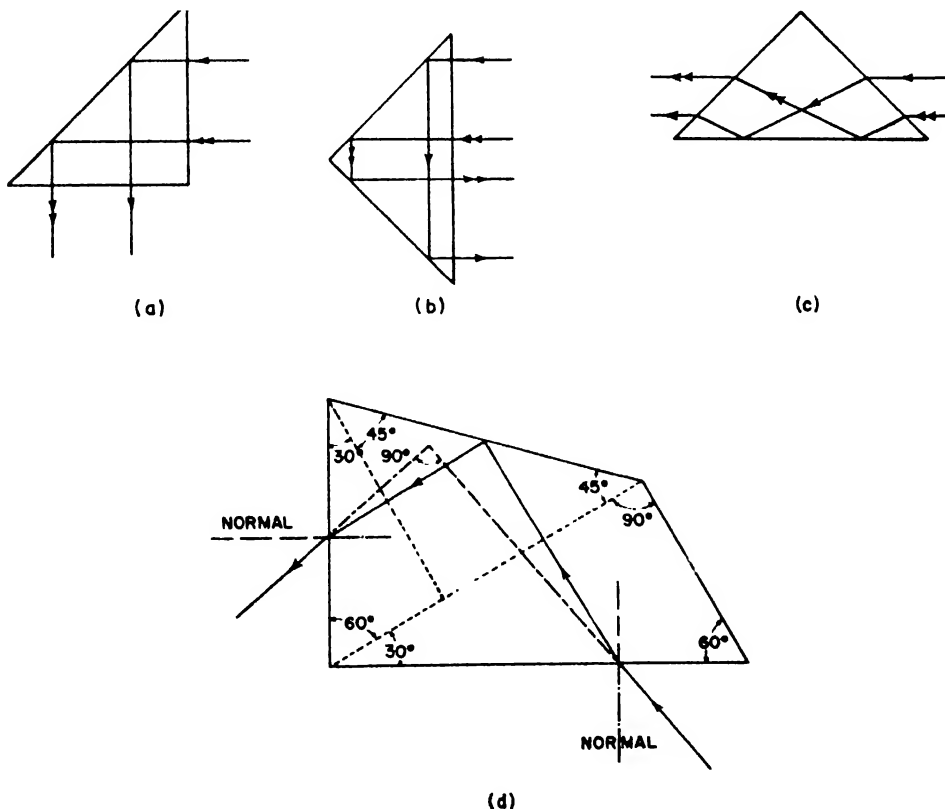


FIG. 3. a, b and c show three methods of using the 45° prism. (d) is a constant deviation prism as used in spectrometers.

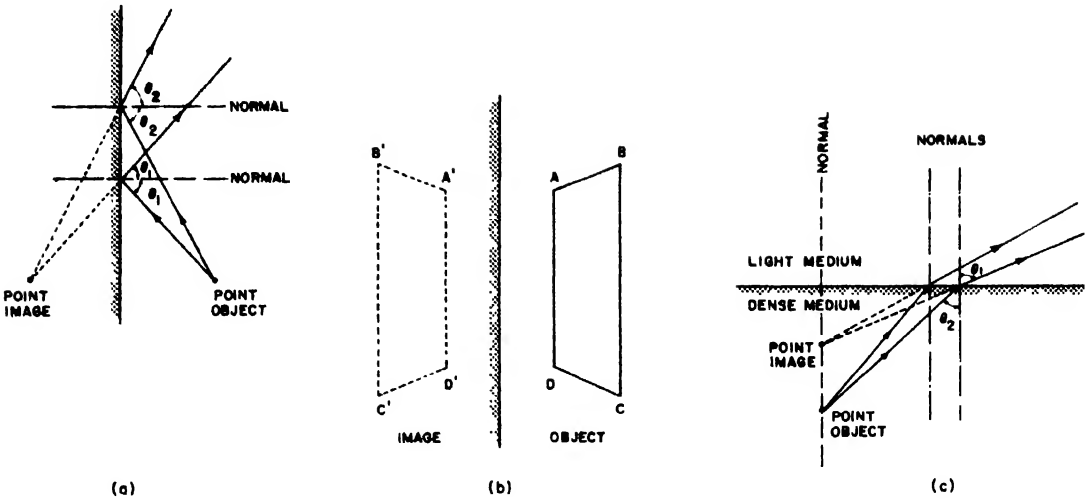


FIG. 4. Image formation by plane surfaces. (a) point object, (b) extended object (c) image by diffraction of a point object.

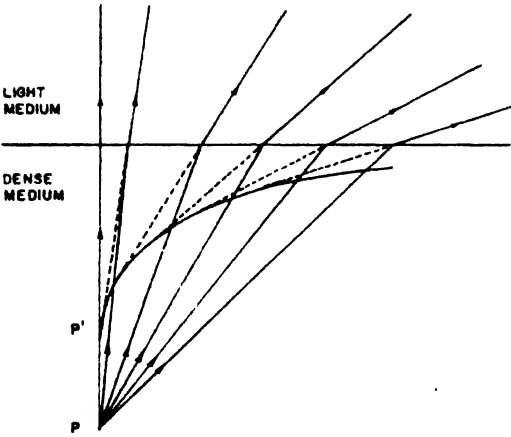


FIG. 5 (a). Formation of caustic by refraction. P, point-object P', position of point-image for small angles of incidence

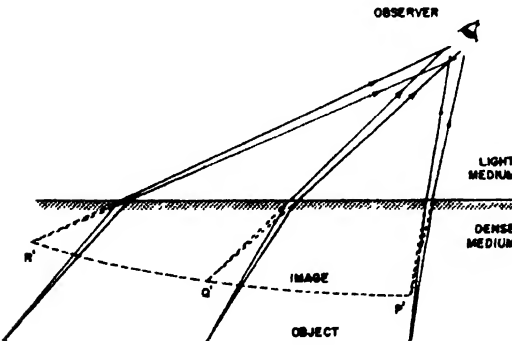


FIG. 5 (b). Image by refraction of an extended object.

of curvature. In each Figure, O is the center of curvature, OP is a radius, and O' is the "pole" of the mirror. A ray parallel to the axis is reflected to a point F, the focus. In Fig. 7(a), F is in front of the mirror; in Fig. 7(b) it is behind, so that the reflected ray *appears* to come from F. For small apertures $O'F \approx PF$ and the position of F is independent of O within this limitation. This gives for O'F, the focal length $f = \frac{r}{2}$.

Image Formation. For an object of small dimensions, the image can be obtained by drawing two rays only: (a) a ray through O—this strikes the mirror normally and is returned along the same path; (b) a ray parallel to the axis—this is reflected through the focal point F. In Fig. 8, AB is the object and A'B' the image. In Fig. 8(a) the image is real, in Fig. 8(b) it is virtual. An image is real when the reflected rays actually pass through points in it; it is virtual if they have to be traced back, so that they only appear to come from the image.

Conjugate Foci. In Fig. 9(a) a point object A produces a point image A'. If A and A' are interchanged, the rays take identical paths but the arrows are reversed, A and A' are called conjugate foci. In Fig. 9(b) the image A' in the convex mirror is virtual. The relationship between the angles is as follows:—

$$\gamma = \theta + \beta, \quad \beta = \theta + \alpha; \quad \therefore \alpha - \gamma = 2\beta$$

For small apertures

$$\alpha = PO'/O'A, \quad \beta = PO'/O'O, \quad \gamma = PO'/O'A'$$

$$\therefore \frac{1}{v} + \frac{1}{u} = \frac{2}{r} = \frac{1}{f}$$

This is a purely quantitative relationship. For general application it is necessary to adopt a convention regarding signs. Distances are always measured from the mirror, and the direction of the incident light is reckoned positive. As an example let $u = 8$ and $r = 6$.

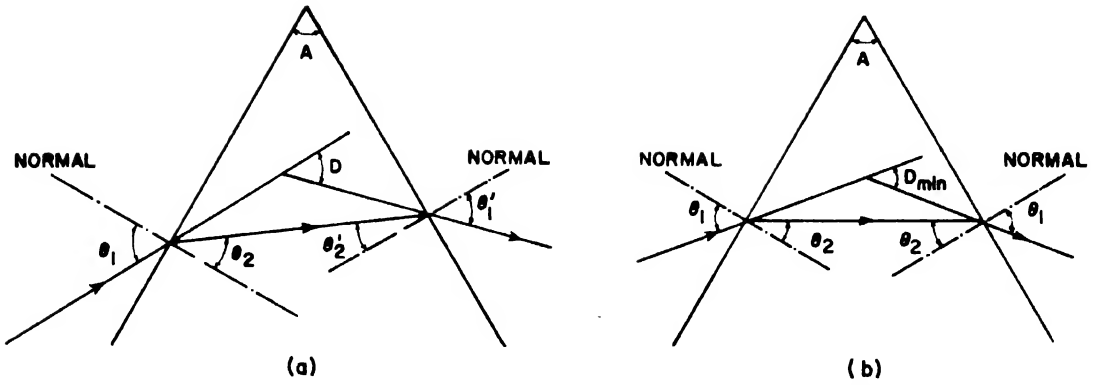


FIG. 6. Deviation by prism. (a) general case, (b) condition for minimum deviation.

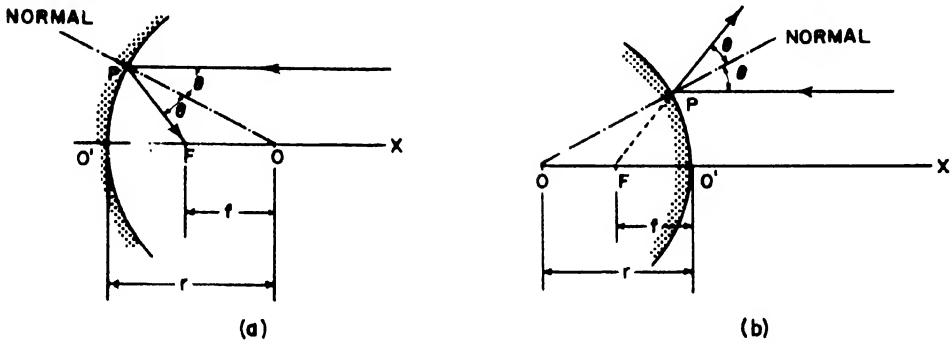


FIG. 7. Reflection at spherical mirrors (a) Concave (b) Convex.

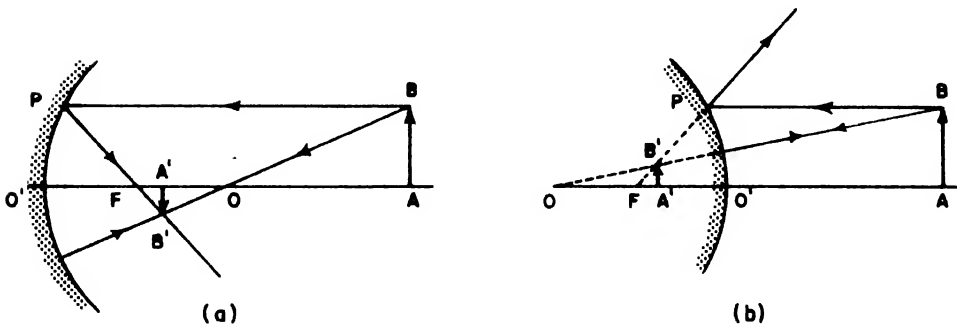


FIG. 8. Image formation in spherical mirrors. (a) Concave (b) Convex.

Concave Mirror	Convex Mirror
u and r negative $u = -8, r = -6, f = -3$ $\therefore \frac{1}{v} = -\frac{1}{3} - \left(-\frac{1}{8}\right) = -\frac{1}{4.8}$ $v = -4.8$	u negative and r positive $u = -8, r = +6, f = +3$ $\therefore \frac{1}{v} = -\frac{1}{3} - \left(-\frac{1}{8}\right) = -\frac{11}{24}$ $v = +2.18m$

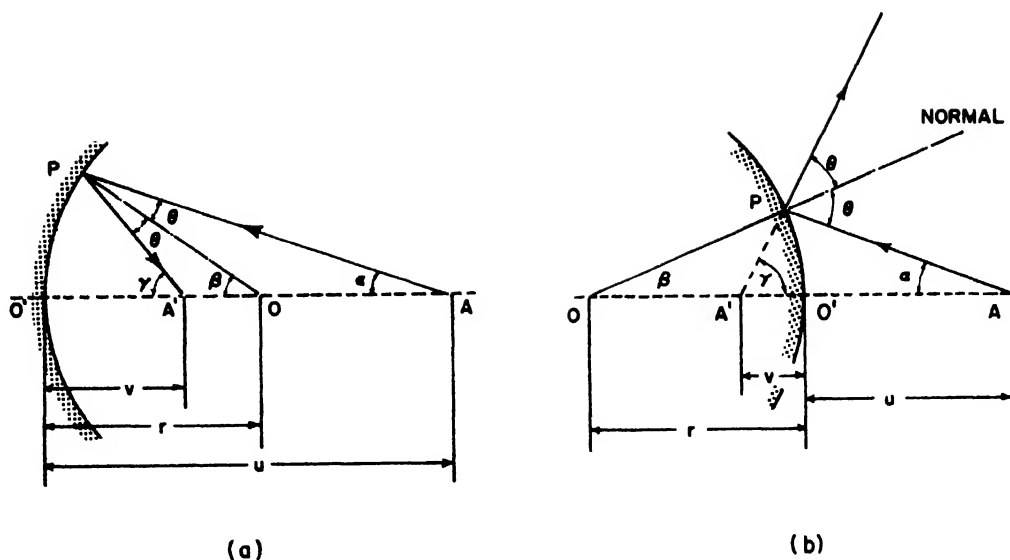


FIG. 9. Conjugate foci.

The magnification

$$m = \frac{A'B'}{AB} = \frac{v}{u}$$

This is positive if v and u have the same sign as with the real inverted image of a real object, Fig. 8(a). It is negative if u and v are of opposite sign as with the virtual erect image of a real object, Fig. 8(b).

The axial magnification of an object of finite axial dimension is given by differentiating the above equation.

$$-\frac{dv}{v^2} - \frac{du}{u^2} = 0, \quad \therefore \frac{dv}{du} = -\frac{v^2}{u^2}$$

This shows that if an object is moved towards the mirror a distance du , the image will move $dv = -v^2 du/u^2$. Thus the image will move more rapidly when u is small.

Parabolic Mirror. For large apertures for which the angle β is not small, the rays of a parallel axial beam are not all brought to a focus at F but form the envelope of a cusp, as in Fig. 10(a). A parabolic mirror brings all the rays of such a beam to the same focus, [see Fig. 10(b)].

Refraction at Spherical Surfaces. The geometry of Fig. 11(a) gives for a surface of small aperture

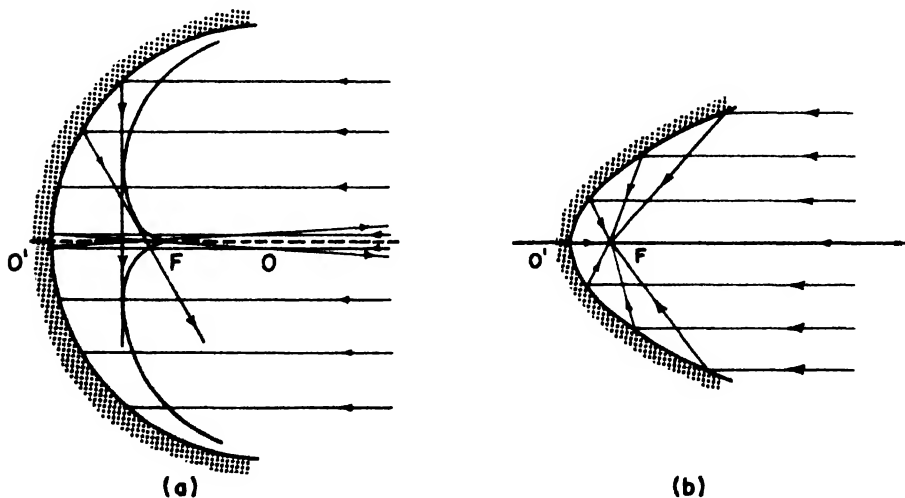


FIG. 10. Reflection by mirrors of wide aperture (a) Spherical, (b) parabolic.

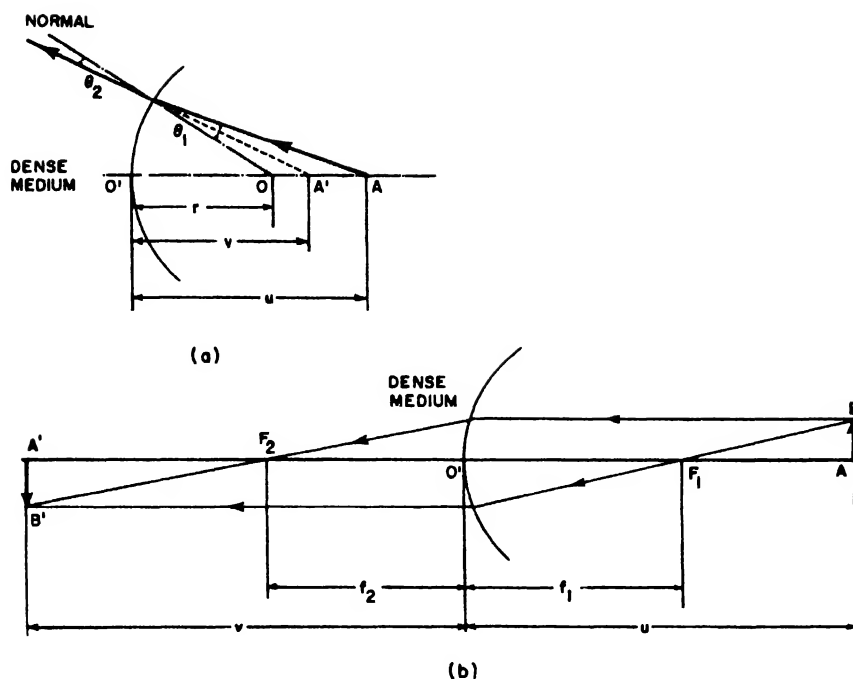


FIG. 11. Refraction at spherical surface of separation between two media.

If we put $v = \infty$ so that the rays are parallel after refraction and put f_1 for the corresponding value of u

$$\frac{n}{r} - \frac{1}{f_1} = \frac{1}{v}$$

If we put $u = \infty$, so that the incident rays are parallel, the corresponding value of v is given by

$$\frac{n}{r} - \frac{1}{f_2} = \frac{1}{v} \quad \therefore \frac{n}{f_2} - \frac{1}{f_1} = 0$$

For an object AB, Fig. 11(b), we obtain the position of the image by drawing two rays from B; the first parallel to the axis, thereby passing through F_2 after refraction; the second passing through F_1 thereby becoming parallel to the axis after refraction.

$$m = \frac{f_1}{u - f_1} \quad \text{or} \quad \frac{v}{f_2}$$

Lenses. Figure 12 shows three media separated by spherical surfaces. Considering each surface in turn

$$\frac{n_2}{n_1} \cdot \frac{1}{w} - \frac{1}{u} = \left(\frac{n_2}{n_1} - 1 \right) / r_1$$

$$\therefore \frac{n_2}{w} - \frac{n_1}{u} = \frac{n_2 - n_1}{r_1}$$

Similarly

$$\frac{n_3}{v} - \frac{n_2}{u} = \frac{n_3 - n_2}{r_2}$$

adding,

$$\frac{n_3}{v} - \frac{n_1}{u} = \frac{n_3 - n_2}{r_2} + \frac{n_2 - n_1}{r_1}$$

In the case of a lens, $n_3 = n_1$, giving

$$\frac{1}{v} - \frac{1}{u} = \left(\frac{n_2}{n_1} - 1 \right) \left(\frac{1}{r_1} - \frac{1}{r_2} \right)$$

with air on either side of the central section, $n_1 = n_3 = 1$, and we can put $n_2 = n$.

$$\therefore \frac{1}{v} - \frac{1}{u} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right)$$

If we put $u = \infty$ then $v = f$, the focal length,

$$\therefore \frac{1}{f} = (n - 1) \left(\frac{1}{r_1} - \frac{1}{r_2} \right)^* \quad ; \quad f = \frac{r_1 r_2}{(n - 1)(r_2 - r_1)}$$

$$\text{Lateral magnification } m = \frac{v}{u} \quad \text{or} \quad \frac{f}{u - f}$$

$$\text{Axial magnification } \frac{dv}{du} = - \frac{v^2}{u^2}$$

$$\text{Example } r_1 = 10, \quad r_2 = 15, \quad \mu = 1.6$$

* Some authors give this equation with a + sign in the second bracket. In such a case it is necessary to use different signs for convex and concave surfaces, a confusion which is avoided by the simple convention adopted in this article.

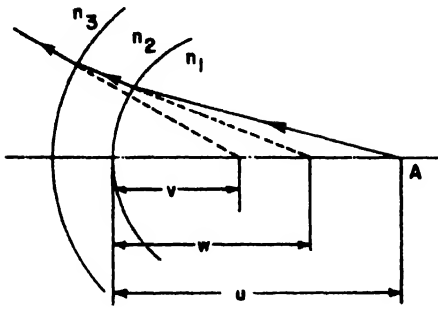


FIG. 12. Lens.

With the convex lenses the focus is on the far side of the lens from the source; with the concave lenses it is on the same side as the source. With plano-convex and plano-concave lenses, $1/r$ for the plane surface is zero.

Defects in Mirrors and Lenses. An aperture which is large compared with the focal length introduces errors in both mirrors and lenses. One of these is the axial spread of the focus as shown in Fig. 10(a). The amount of spread for a mirror of aperture 2θ is

$$\frac{r}{2} \left(\frac{1}{\cos \theta} - 1 \right)$$

Astigmatism results when the object is situated off the axis. The reflected pencil is brought to a focus at a point I, also off the axis, and is spread into an axial line Q_1Q_2 , as in Fig. 13(a). If the mirror is turned through a small angle, the point I describes a line called the first focal line. Q_1Q_2 is the second focal line and is perpendicular to the other. In between there is a region where the reflected pencil is the nearest approach to a circle; this is called the circle of least confusion.

Figure 13(b) shows longitudinal spherical aberration in a lens; rays very close to the axis pass through the focal point as defined for a thin lens, while increasing divergence brings the focus nearer to the lens. Figure 13(c) shows chromatic aberration—the focal length, even for a thin lens, being a function of the wavelength. This is because n varies with the wavelength and increases as the wavelength decreases. Thus the focal length for violet is less than for red.

Curvature of the Field. The image of a plane object lies, in general, on a slightly curved field (Fig. 14), and this combines with the effects of astigmatism. Pincushion distortion results when a lens is used as a magnifying glass; barrel distortion results when an object is viewed through a lens at some distance from the eye.

Systems of Lenses. The inverse of the focal length of a lens is called the power of the lens. If

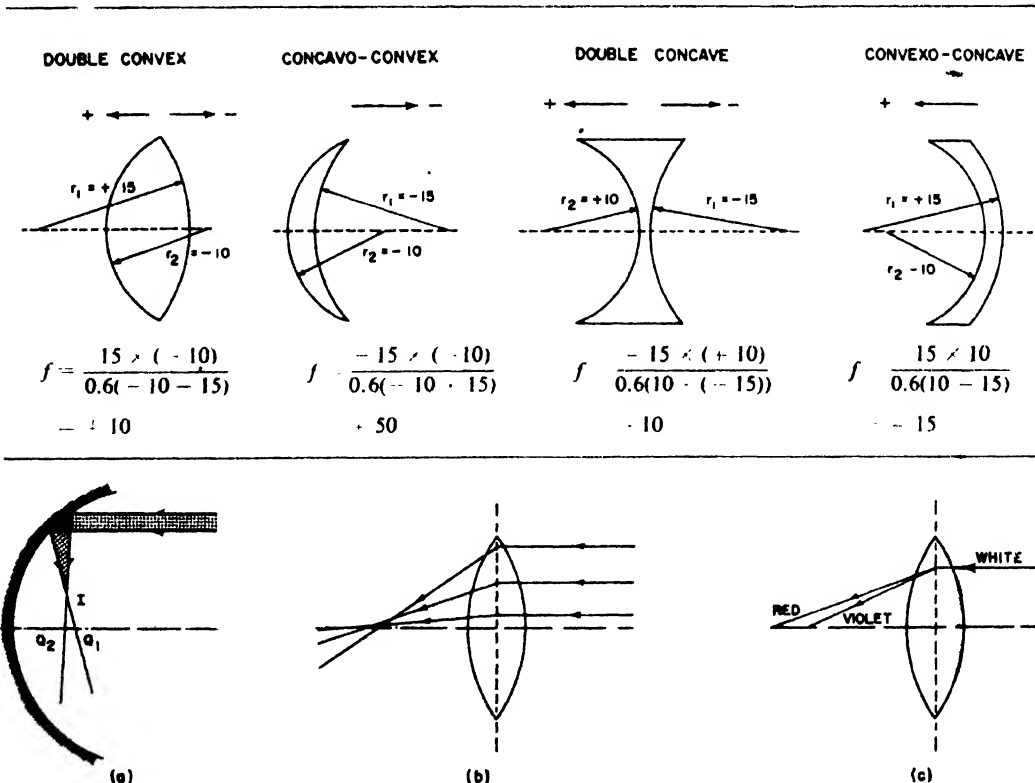


FIG. 13. Image defects. (a) Astigmatism, (b) Longitudinal spherical aberration, (c) Chromatic aberration.

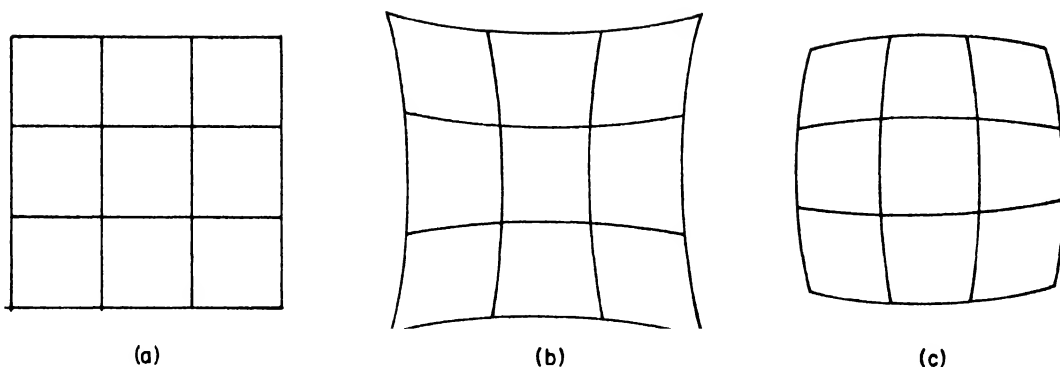


FIG. 14. Image distortion. (a) Object, (b) pincushion distortion, (c) barrel distortion.

the unit of length is the centimeter, the unit of power is the diopter:

$$D = \frac{1}{f}$$

If a number of thin lenses are in contact, the power of the combination is the *algebraic sum* of the individual powers. As an example, consider two thin lenses for which $f_1 = 10$ and $f_2 = -15$ cm.

$$D_1 = \frac{1}{10} = 0.1 \text{ diopter}; D_2 = -\frac{1}{15} = -0.0667 \text{ diopter}$$

$$D = 0.1 - 0.0667 = +0.0333$$

$$\therefore f = +30 \text{ cm}$$

If two coaxial lenses are separated a distance d , then the focal length of the combination is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}$$

For a minimum longitudinal spherical aberration, the distance d should be

$$d = f_1 - f_2$$

H. COTTON

Cross-references: ABERRATION THEORY; LENS; LIGHT; OPTICS, PHYSICAL; REFLECTION; REFRACTION.

OPTICS, PHYSICAL

In this branch of the subject of optics, we deal with phenomena connected with the nature of light itself. We assume that it is a transverse wave motion and that the underlying principles are embodied in the electromagnetic theory conceived by James Clerk Maxwell (1831–1879). A very brief outline of the major phenomena will be all that can be attempted in this article.

An excellent prerequisite to an understanding of the progress of light is Huygens' principle. Christian Huygens (1629–1695), a contemporary of Sir Isaac Newton, doubtless got the idea of

wave propagation by observing the progress of ripples on a Dutch canal. The famous principle, applicable to surface waves on a liquid, to sound waves and to light waves, is that *each point of a wave front may be regarded as a source of new waves*. In Fig. 1, the wavefront W may be regarded as the envelope of an infinite number of wavelets forming from the preceding wavefront W' , and W' may be likewise formed from W .

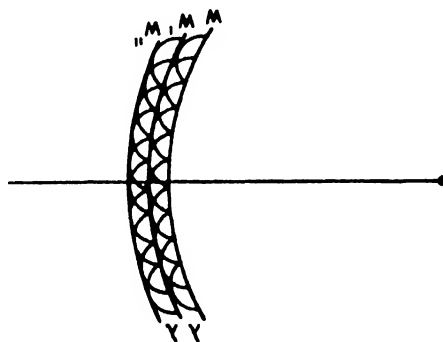


FIG. 1

Interference. A very important special case of the superposition of waves occurs when two identical wave trains W_1 and W_2 , of wavelength λ (Fig. 2), start out in phase from a source S and travel to a point of crossing S' —the one train going directly to S' , the other train going by way of a reflecting (or refracting) device M , and thus traversing a greater distance. The "distance" in each case is known as the *optical path*. What will be the effect at S' due to the superposition of the two coherent wave trains? The path difference taken by W_2 and W_1 is arranged experimentally so that it is a whole number of half wavelengths, or

$$(SM + MS') - SS' = n \frac{\lambda}{2}$$

- (i) when n is odd, there is *destructive interference*
- (ii) when n is even, there is *constructive interference*.

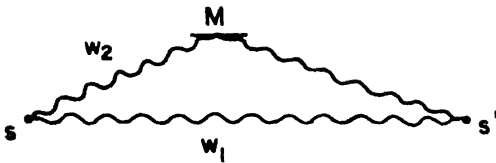


FIG. 2

In (i) the amplitudes of the two waves will cancel, since there is a phase difference of 180° , and light plus light will produce darkness at the point S' . In (ii) the amplitudes of the two waves will add, and there will be a maximum intensity at the point S' . It is assumed that the vibrations of the two wave trains are in the same *plane* in each instance. If they are not, a new effect is observed. For points near S' on either side, there will be gradations between zero and a maximum intensity. In the hands of outstanding experimenters such as Albert A. Michelson (1852-1931) the phenomena of interference have been put to great practical use in the measurement of extremely minute distances and angles (see INTERFERENCE AND INTERFEROMETRY).

It was the behavior of light during interference that enabled Thomas Young (1773-1829) to enunciate the wave theory of light and thus to upset the corpuscular theory which had been held by Newton. Material particles would not show interference, but wave trains—under the carefully specified conditions—would do so. As Young's crucial experiment is performed today, a train of plane waves W of wavelength λ is incident upon a screen C_1 (normal to the plane of the diagram, Fig. 3) which has cut in it a narrow slit S that serves as a line source of an emerging train of semicylindrical waves W' , in accord

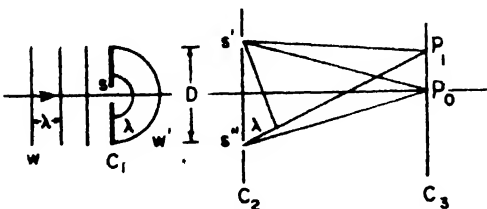


FIG. 3

with Huygens' principle. Parallel to C_1 and relatively far away is a second screen C_2 which has cut in it two narrow slits S' and S'' , separated by a distance D and equally disposed on the two sides of the normal to C_1 at S . By Huygens' principle, S' and S'' , being parts of the same wave front, will send out wave trains that are coherent. If the normal through S cuts a third screen C_3 , that is parallel to the other two, at P_0 ; then at this point there will be zero path difference between it and the two slits S' and S'' . A bright line of light will be seen here, due to constructive interference. For two other points, P_1 and P_1' ,

equidistant from P_0 , the path difference $P_1S'' - P_1S'$ will be λ . Here again there will be constructive interference. Upon the screen C_3 there will be seen an interference pattern.

Diffraction. In general, the term diffraction denotes the departure of a train of waves from a straight course. Thus, ocean waves, parallel to the shore outside of an inlet, are seen, when they enter the harbor, to run along the inner shores—diffracted by the passage from the inlet to the harbor. Sound waves in a concert hall are readily heard in the "shadow" of a column that supports a gallery. By the same principle, light waves bend into the geometrical shadow when passing a straight edge. If, in the discussion of interference from two slits (Fig. 3), the widths of the slits were increased so as to include a larger number of "points" on the incident wavefronts, then the effects on the third screen C_3 would be more complex (see DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS).

Dispersion. As used in optics, dispersion implies the separation of a light beam of more than one wavelength into its component parts. The operation can be effected by means of a diffraction grating or by a prism. In the simple case of dispersion by a grating G (Fig. 4), the

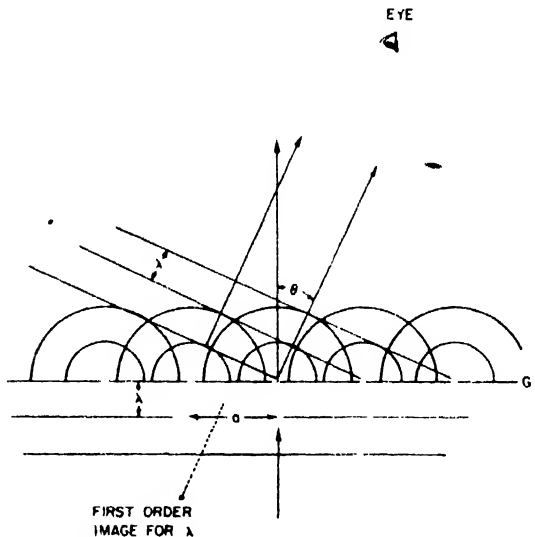


FIG. 4

wavelength λ , the angle of dispersion θ and the distance a between the centers of any two consecutive openings are related by $n\lambda = a \sin \theta$, where n , a positive integer, known as the *order* of the spectrum, is the number of whole wavelengths difference in path between the observer's eye and the centers of any two consecutive openings—assuming that parallel light is incident upon the grating. For dispersion by a prism, the angular spread of spectrum between the limits λ_1 (e.g., violet) and λ_2 (e.g., red) is given by $\alpha \approx A(n_1 - n_2)$ if the angle of the prism A is "small," say 15° or

less, and the prism is very nearly at minimum deviation for the extreme wavelengths. Prismatic dispersion is observed with a spectrometer or a spectroscope. The most celebrated example of prismatic dispersion was Newton's experiment with a prism held in the path of sunlight. The measure of dispersion depends upon a definite prism of given material and refracting angle, together with the values of the wavelengths involved. The *dispersive power* of the prism

material is defined by $p = \frac{n_1 - n_2}{n_1 - 1}$ between the limits 1 and 2, the center of the range being at some intermediate wavelength λ_3 . The angular dispersion of a given prism, or rate of change of deviation D with wavelength, may be regarded as the product of the quantity dD/dn , calculable by geometrical optics, and the quantity $dn/d\lambda$ (a property of the material). The last may be found from the *normal dispersion curve* (Fig. 5) plotted from data taken with a spectrometer using a prism of the given material and employing light

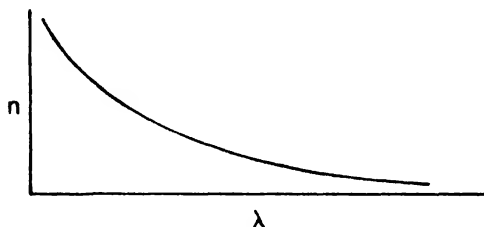


FIG. 5

sources emitting known wavelengths. The angular dispersion is found from a consideration of the equation

$$\frac{dD}{d\lambda} = \frac{dD}{dn} \cdot \frac{dn}{d\lambda}$$

In normal dispersion, the index of refraction decreases with wavelength; i.e., the velocity of light in the material increases with wavelength, and the rate of change of index with wavelength decreases with greater wavelengths. The phenomenon of dispersion is of prime importance in the science of spectroscopy but is a decided liability in the design of lenses for optical instruments, since any one lens by itself will not focus all colors at the same point. Exceptions to the general trend of index variation with wavelength are very common for certain substances throughout particular wavelength bands. The topic of anomalous dispersion is to be treated under the subdivision absorption spectroscopy (see ABSORPTION SPECTRA).

Polarization. Although the wave theory has been very satisfactory in "explaining" the phenomena of interference and diffraction, the crucial test for the necessity of that theory was not presented. What sort of waves are they? Are they longitudinal (as in sound), or circular (as in or on a water surface), or are they transverse (as

along a violin string)? Strong experimental evidence, as well as the electromagnetic theory of light itself, point to the transverse type of wave form for light—as well as for radiant heat, radio waves, and the other regions of the so-called *electromagnetic spectrum*. In Fig. 6(a) imagine that a beam of natural light is emerging from the paper at O and that the "light vectors" can take in turn any possible direction making an angle with the x-axis. If, now, a very finely ruled grill PP' ("polarizer") is placed in the beam as in Fig. 6(b), only those vibrations, or components of vibrations, which are parallel to PP' will be transmitted. The light that emerges will be *plane polarized*. Now let a similar grill AA' ("analyzer") Fig. 6(c) be placed in this transmitted beam, with an angle θ between the directions PP' and AA'.

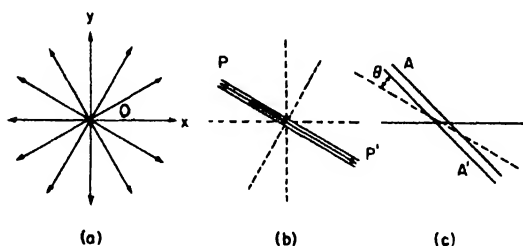


FIG. 6

Then only those vibrations, or components of vibrations, emerging from PP' which are parallel to AA' will be transmitted by the latter. The incident light (from PP') has therefore been "analyzed." An everyday example of plane polarization was first observed by Etienne Louis Malus (1775-1812) through a chance "analysis" of sunlight reflected from a distant window. The Malus effect may be understood from Fig. 7. AO is a beam of natural light in the plane of the paper, incident upon the surface MM' at an angle i . The plane of the paper is the plane of incidence and MM' is perpendicular to it. The beam of natural light has as many light vector components at right angles to the plane of incidence (· · · · ·) as there are in that plane

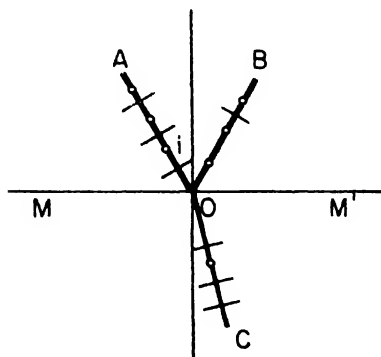


FIG. 7

(// //). Malus found that the reflected beam (OB) is partially plane polarized, which means that there are in it more of the (· · · ·) components than there are of the (// //) components. The situation is reversed in the refracted beam (OC). Several years later, David Brewster showed that if the reflected and refracted beams are at right angles to each other, the polarization of the former is complete and that the tangent of the then angle of incidence is the index of refraction of the material (for that wavelength).

Optical Activity. It is not always true that a beam of plane polarized light maintains its vibrational direction as it proceeds into a transparent substance. Many materials actually rotate the plane of vibration by an amount that depends directly upon the distance traversed. *Optical activity* is exhibited by quartz, sodium chlorate, turpentine, sugar crystals (even if dissolved in water), etc. In the case of a solution, the rotation is also proportional to the concentration, and it may be clockwise or counter-clockwise. The wavelength of the light and the temperature are also factors in the total rotation observed. The apparatus used in the measurement of the angular displacement of the incident beam is a *polarimeter*.

Double Refraction. Certain materials, such as quartz, mica, Iceland spar and tourmaline, are able to refract a beam of light into two portions which, in general, travel thereafter in different directions. The phenomenon of *double refraction* was first described in 1669 by Erasmus Bartholinus. In Fig. 8 we imagine a beam of natural light of a

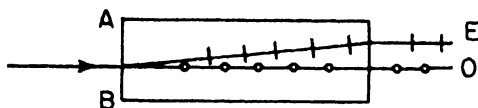


FIG. 8

given wavelength to fall upon the face AB of a crystal of Iceland spar (a clear form of calcium carbonate). One part, after passing the interface, travels on through the crystal in accord with Snell's law, and is known as the *ordinary* beam. The other part of the incident beam, after passing the interface, goes off at an angle with the first part. This is the *extraordinary* beam. If the crystal faces are parallel, the two beams will emerge parallel. It was soon found that they are plane-polarized at right angles to each other. The symbols (·) and (//) serve to denote this fact. Each crystal has an optic axis along which optical properties remain constant. If the incident light is normal to, or parallel to, the optic axis, there is no separation of the ordinary and extraordinary beams. Huygens explained double refraction of Iceland spar in terms of concentric spherical and ellipsoidal wavefronts (Fig. 9). The small double-headed arrows denote the vibrational directions of the transverse waves.

Photoelasticity. Many ordinary materials, such as plastics, can be made doubly refracting if they are placed under a stress. This new property will

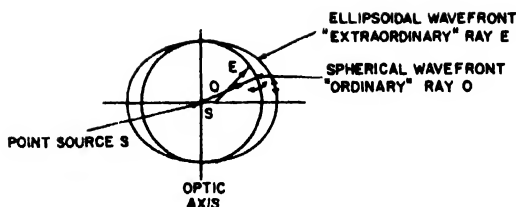


FIG. 9

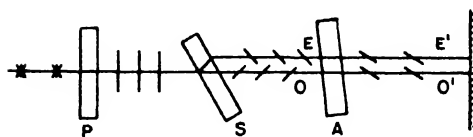


FIG. 10

show up if the specimen is placed between two "crossed" polarizing sheets (Fig. 10). Natural light is plane-polarized by the first sheet (polarizer) P and is then split by the specimen into ordinary and extraordinary beams, O and E. The second polarizing screen (analyzer) A suppresses one component of each beam so that the light vectors O' and E' that emerge from the analyzer are in the same plane and are in the condition to show interference at the screen.

C. HARRISON DWIGHT

Cross-references: DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS; INTERFERENCE AND INTERFEROMETRY; OPTICS, GEOMETRICAL; POLARIZED LIGHT; REFRACTION.

OSCILLOSCOPES

The cathode-ray oscilloscope is an instrument that displays changing events in graphical form on the screen of a cathode-ray tube. The display is usually presented in terms of X-Y coordinates, with time represented horizontally from left to right, and the electrical analog of the event by a vertical displacement, up or down from some zero point (the abscissa and ordinate, respectively) (see ELECTRON OPTICS).

The first known use of the cathode-ray tube to display changing phenomena occurred in 1897, the same year that the tube was used to prove that cathode rays were electrified and would respond to a magnetic field. As the oscilloscope has come into greater use, it has been required to permit more and more accurate measurements of time and amplitude, directly from the display if possible. The engineering efforts to satisfy these requirements have influenced the basic philosophy of the instrument, have initiated developments in circuitry and components as well, and have resulted in sophisticated instruments that are actually small but complex "systems." For example, some "straightforward" oscilloscopes

meant for broad general-purpose laboratory use contain more than 150 vacuum tubes and semiconductor devices, while a sampling oscilloscope will contain nearly 250. Other instruments, designed to measure and automatically print out several parameters of semiconductor devices at the rate of several per second, contain over 600 semiconductors and tubes. Actually, in this last example, the cathode-ray tube has assumed the important but subsidiary function of a monitor for setting up and observing the desired performance of the testing equipment.

The events displayed by an oscilloscope are usually referred to as signals, and the parameters of the signal which it is desired to observe determine the characteristics the instrument must have. The design of an instrument, therefore, should reflect the desired range of signals it is expected to display. An instrument to deal with a narrow range of uniformly repetitive signals will be quite simple and will exhibit a different design philosophy than one expected to deal with arbitrary signals. These latter impose the most stringent design requirements, as there is no way of anticipating the time of signal occurrence, how soon it will occur again (if ever), duration, and its character or "form" (which determines its content).

An oscilloscope consists of three or four basic elements: cathode-ray tube, time-base generator, vertical deflection channel and power supply (which is quite important, as it sets the limits to both long-term and short-term stability of performance of the other elements). (Many would say that the signal pick-up is a basic element since it determines whether the instrument has a "good" signal to display.)

The case of a signal consisting of a single arbitrary event illustrates the relationship between these basic elements. Such a signal, or one whose duration is very short compared with the time until its next occurrence, presents a special problem whose solution permits viewing almost any other kind of signal. The problem can best be stated in the form of a question: How can all of such a signal be seen when the chain of complex events which permit it to be displayed must wait until its occurrence to be initiated?

The time relationships involved require that the signal not appear on the vertical deflection plates of the cathode-ray tube until the time base has been impressed on the horizontal deflection plates. Clearly, the signal must be "held up," stored, or somehow delayed until the sweep generator is well started and the electron beam in the cathode-ray tube is flowing. Normally this delay is performed in the vertical amplifier by a transmission line which requires a large amount of time for the signal to traverse compared to the onset of the signal and the starting time of the sweep.

Signal amplifiers in the cathode-ray oscilloscope are usually required to respond accurately to both the dc level at which a signal may occur and the signal itself, no matter how transient it

may be. At the present time, signals between 10^{-11} and 10^{-12} second are of interest, and present-day vacuum tube amplifier techniques are simply incapable of handling such a range of signals.

This problem has been approached in several ways, using vacuum tubes, semiconductors and a combination of the two technologies.

The vacuum-tube approaches reflect the fact that electrical currents and fields take time to move from place to place—whether through the space in a vacuum tube or along a conductor from one component to another. This fact makes it possible to space electrical elements physically so that they will influence the same electron or signal at different times. Consequently, we can apply the same deflecting field to the same electron or amplify the same signal a number of times at successive intervals. Sizeable improvements in bandwidth at a given deflection sensitivity in a cathode ray tube, or in gain at a given bandwidth in an amplifier, are possibly by these techniques. Devices designed in this way are often referred to as "distributed deflection plate" cathode-ray tubes or "distributed" amplifiers. (Traveling wave amplifiers, cyclotrons, klystrons and magnetrons all employ the principle of using the same field to affect the same electron at successive instants of time.)

Semiconductor devices in the proper physical form are capable of responding in exceedingly short periods of time—a few picoseconds (10^{-12} second), in fact. This capability, coupled with the recently developed microcircuitry or deposited film technology, should make some interesting and useful devices possible.

Another way in which an apparent increase in bandwidth can be obtained is by the sampling method. It is possible to generate very short voltage pulses in a very precise manner, both as to time of occurrence and form. If such a pulse is combined with a repetitively recurring signal at successively later instants (in relation to some arbitrary "point" on the signal), a series of voltage pulses is developed whose amplitude varies in the same manner as that of the signal in question, but whose duration is expanded by as much as is desired. In other words a signal of very short duration can be expanded into one of long duration, which can be handled by conventional amplifier techniques. An effective increase in bandwidth of 100 times or more can be obtained by this method (see PULSE GENERATION).

The development of deposited film technology makes it possible to obtain extremely wide-range performance never before attainable with vacuum tube or even conventional transistor technology. These new techniques hold a great deal of promise for obtaining increased performance in the oscilloscope art.

In at least one respect, the oscilloscope differs from most other instruments; it is expected to display the change in some phenomenon without distortion in either time or magnitude. In contrast with most other instruments, it does not deal with only one aspect of a phenomenon, such as an

average value, a peak value, the time between two events, the number of events, etc. In these cases, the attributes of no concern can be distorted if necessary to facilitate measuring the desired attribute. The oscilloscope, however, must display the "way" in which something changes, and it should "process the information" between the receiving of it and the displaying of it in an essentially passive manner.

Consequently, the use of the oscilloscope implies that it "displays" precisely what is put into it (otherwise it would not be very useful!). This is not absolutely possible in the general case because certain signals whose characteristics fall near or outside the range of the oscilloscope amplifier are distorted to some degree. Such distortion is inevitable. The problem is to determine which kind of distortion is acceptable.

Ordinarily, bandwidth has been considered to be the principal criterion of performance of an amplifier. However, when an amplifier is handling arbitrary signals, it is nearly impossible to adjust it to give undistorted response over the whole range of performance by using conventional bandwidth-testing techniques. The most effective method that has been developed to adjust an amplifier for uniform undistorted response is by the use of step functions or square waves. A step function is a "no-time" transition from one dc voltage level to another. A square wave is simply a series of step functions "alternating" between two dc levels of voltage in a regular manner. The value of a square wave is that its quality can be evaluated by eye with a precision unmatched by most conventional, easily used bandwidth-testing techniques. Once the criteria of distortion have been agreed upon, evaluations can be made which are limited only by the skill and experience of the viewer.

Several kinds of distortion of a step function may occur in an amplifier. For example, it can be adjusted to give the shortest possible risetime, since the step function rises in "no time." Although such an adjustment gives the best possible reproduction of the actual transition, it also causes the signal to "overshoot," adding high frequencies to the display that were not present in the original signal. Conversely, the amplifier can be adjusted so that the dc levels between the transition are completely undisturbed and the result is a gradual transition from one voltage level to the other which subtracts high frequencies from the signal. However, there is an optimum region of performance where the transition takes place as rapidly as is possible under the circuit conditions but with no overshoot and a minimum of detectable distortion in the stable voltage levels following the transition. In this situation, the circuit itself determines the response to a step function, the deteriorated portion of the step function can be computed from the observed risetime and the known amplifier risetime, and most important, there is never any question whether a particular element of the display is due to the amplifier or is actually present in the signal. If something appears in the display, it is

present to at least that degree in the signal. It may be present in the signal to a greater degree than the display indicates, in which case responses of about one-fifth of the risetime of the amplifier can be estimated to a reasonable accuracy.

The problem of determining fidelity of response is a never-ending one. The testing square wave must be "square" enough that it does not contribute any distortion to the ultimate display. However, only by having a measuring system with capabilities much beyond the characteristics of the testing square wave can we be reasonably confident of the quality of this signal. But we can only test this latter measuring system for fidelity of response by having another testing square wave of known fidelity so that any distortion in the amplifier will be apparent to the user. The solution to this problem takes on the character of attempting to lift oneself by one's own bootstraps and, of course, is never satisfactorily resolved.

JACK E. DAY

References

- MacGregor-Morris, J. T., and Mines, R. J., *J. Inst. Elec. Engrs.*, 63, 1056 (1925).
- Kuehni, H. P., and Ramo, S., *Elec. Eng.*, 721 (1937).
- Day, J. E., "Recent Developments in the Cathode-Ray Oscilloscope," *Advan. Electron Electron Phys.*, 10, (1958).
- Noel, D. R., and Susskind, C., *Elec. Ind.*, 92-98 (August 1961).
- Kobbe, John, "The Sophisticated Oscilloscope," *Industrial Research* (March 1964).

Cross-references: ELECTRON OPTICS, PULSE GENERATION.

OSMOSIS

If, into the bottom of a jar containing water, a solution of cane sugar is introduced with care so as to avoid mixing, not only will the molecules of cane sugar diffuse into the water but the molecules of water will diffuse into the sugar solution. These processes will go on until the concentration of sugar, and of water, is the same throughout.

If the solution is placed in a container, whose walls are relatively impermeable to the sugar while being permeable to the water, and the container is placed in water, the water will pass from the outside into the container. The term osmosis is usually restricted to the passage of water. If the influx of the water results in an overflow of solution to somewhere other than the surrounding water this overflow will continue until all the sugar is removed from the container. If the container is closed, water will continue to enter until there is sufficient stress in the stretched walls to cause a pressure on the solution inside; this will eventually stop the influx. Of course, if the walls of the container are not completely impermeable to sugar, then the sugar will be escaping into the water outside the container and this will go on until the concentration of sugar is the same out-

side and inside. If the walls of the container were impermeable to water but permeable to solute, the latter would escape. The cause of this osmosis, this "pushing," of water into the solution is that the tendency of the water molecules to escape from the pure water is greater than that of the water molecules in the solution. Consider water in contact with a limited volume of air. Of those molecules of water striking the surface some will have sufficient energy to escape into the air and this escape will result in net loss to the air which will continue until the concentration of water vapor molecules there is such that the rate of escape from the air (into the water) equals the rate of escape from the water (into the air). If the volume of the air space is fixed, the pressure will rise. Just as the temperature of all bodies is the same when they are in thermal equilibrium, although their heat content per unit volume varies with their specific heat, so the escaping tendency of the water is the same in all systems when they are in aqueous equilibrium, whether the system is pure water, solution, gas phase, wettable solid, etc. The same concept can be applied to any substance, say mercury in pure mercury, in air containing mercury vapor, and in an amalgam with another metal such as zinc. The term osmosis is usually restricted to the passage of water from a solution where the escaping tendency is higher to a solution where it is lower. Moreover it is usually restricted to the passage through a solid or liquid barrier which prevents the solutions from rapidly mixing. It is not used for the passage of water in the form of vapor through the air from a dilute solution to a stronger solution in the same confined space, although the process is fundamentally the same. It is sometimes restricted to the case where the barrier is semipermeable, that is, lets through water but not solute.

The escaping tendency of water is lowered by the addition of a solute. If the molecules of the solute have no other effect than to reduce the number of molecules of water in unit volume, then the escaping tendency of the water will be reduced proportionately to the reduction in the mole fraction of water, N_1 , the ratio of the moles of water to the sum of the moles of water and solute. Such is a "perfect" solution. If, however, there is some attraction between the solute and water molecules, a smaller fraction of the latter will have energy sufficient to escape—a "non-perfect" solution. The escaping tendency is increased by pressure. Hence a solution in which the water has lower escaping tendency than it has in pure water at the same pressure, P^0 , can be brought to water equilibrium by a sufficient increase in the pressure on the solution to a value P . This sufficient increase, $P - P^0$, is the osmotic pressure of the solution. In general we cannot state $P - P^0$, the osmotic pressure, knowing only N_2 , the ratio of moles of solute to the sum of the moles of water plus solute, the mole fraction of solute, ($N_2 = 1 - N_1$).

What we can say is, that if in a solution with a mole fraction N_2 of solute under a pressure P the water has the same escaping tendency as it has in

pure water at the same temperature and at a pressure P^0 , then $dP/dN_2 = A/B$ where dP/dN_2 is the increase of P relative to increase of N_2 to keep the escaping tendency unchanged; A is the decrease of escaping tendency relative to increase of N_2 when P is unchanged; and B is the increase in escaping tendency relative to increase in P when N_2 is unchanged. For dilute solutions A/B approximates to RT/V_1 and so $P - P^0$ approximates to N_2RT/V_1 where V_1 is the volume of one mole of water, R is the constant $82.07 \text{ cm}^3 \text{ atm/deg}$, and T is the absolute temperature. For very dilute solutions N_2/V_1 approaches n_2/V , the number of moles of solute in a volume V of solution and $P - P^0 = n_2RT/V$ (van't Hoff's equation). This gives an osmotic pressure of 1 atm for one mole of solute in 22.4 liters at 0°C . There is a departure from both these equations for stronger solutions. The fact that one mole of a perfect gas in 22.4 liters at 0°C exerts a pressure of 1 atm, coupled with the above, has led some to say that the osmotic pressure is the bombardment pressure of the solute molecules. It is correct to say that for very dilute solutions the osmotic pressure of a solution is equal in magnitude to the pressure the solute molecules would exert if they were alone in the same volume and behaved as a perfect gas, but that is another matter.

To measure the osmotic pressure a semipermeable membrane must be prepared which itself can stand sufficient pressure, or it must be deposited in the walls of a porous pot so that the pressure can be sustained. With the solution being inside and water out, pressure is applied to the former until there is no net movement of water.

Observations by Berkeley and Hartley showed that for 3.393 gms of cane sugar per 100 gms H_2O , the osmotic pressure at 0°C is 2.23 atm while the van't Hoff equation gives 2.17 atm since $n_2/V = 9.72 \cdot 10^{-5}$. If N_2/V_1 is used instead of n_2/V , the value of 2.22 is obtained. With stronger solutions the measured osmotic pressure exceeds that calculated: with 33.945 gms of sugar, 24.55 atm is the value measured, while van't Hoff's equation gives 18.41 and the other 21.8 atm. The observed value is given if, in calculating N_2/V_1 , it is assumed that each sugar molecule immobilizes five molecules of water.

The solutes in the vacuole of a plant cell are exposed to the inward pressure of the distended cell wall and that of the turgid surrounding cells. Water will pass into the cell vacuole from water outside as long as the total inward pressure on the vacuole falls short of the osmotic pressure of the solution in the vacuole. Passage of water into the vacuole dilutes the contents and lowers the osmotic pressure and increases the inward pressure by distension. The amount by which the inward pressure falls short of the osmotic pressure is called by some the suction pressure.

A substance such as cellulose or gelatin tends to take up water, the tendency decreasing with increase in water content until the stress in the substance causes a sufficient rise in the escaping tendency of the water in the substance. This process, which like osmosis is a movement from

higher to lower escaping tendency, is called imbibition, and the pressure on the substance sufficient to stop the uptake is the imbibitional pressure. Hence, if a plant cell with a cellulose wall, after coming to equilibrium with a solution, is transferred to water, the wall takes up water by imbibition and the vacuole by osmosis. The latter considers only the over-all movement from outside to vacuole and does not consider the movement from cellulose to vacuole, a process which is the reverse of imbibition. A plant cell in equilibrium with a solution having an osmotic pressure of 25 atmos would also be in equilibrium with air about 98 per cent saturated with water vapor. If the cell was transferred to a saturated atmosphere, it would take up water. We lack precise terms for the passage of water from air into the cellulose and into the vacuole. Condensation, which might be used, ranges more widely.

The escaping tendency of water is affected by factors other than concentration of solute and pressure. Increase of temperature increases escaping tendency. This is a complex problem involving not only transfer of water but also of heat. To a minor extent, the passage of water from pure water to a solution involves a heat transfer.

For many naturally occurring membranes which are not completely semipermeable, i.e., they let solute molecules through slowly, electro-osmosis is important. If the membrane tends to lose negative charges to, or take negative charges from, water or solutions, then the water molecules, in the pores of the membrane, will tend to take on an opposite charge to the membrane. If there is a gradient of electric potential across the membrane, the charged water will move in the appropriate direction. If the potential difference

is established by the use of electrodes, this is electro-osmosis.

With some membranes, particularly those containing protein, water may pass from a dilute solution on one side to water on the other—negative osmosis. Under such conditions if the solution was acid the membrane would be positively charged through the uptake of H^+ ions, leaving the water in the pores negative. The greater mobility of the H^+ ions relative to the anions will cause the side of the membrane towards the solution to be negative and so drive the negatively charged water in the pores across the membrane in the opposite direction to normal osmosis.

The rate of osmosis depends, not only on the excess of the escaping tendency of the water in the phase from which it moves over that in the phase to which it moves, but also upon the area of surface of interchange and the over-all resistance experienced by the water. The rate of shrinkage of the vacuole of a plant cell when it is placed in a strong solution at first seems surprisingly high. When allowance is made for the fact that the ratio of surface to volume increases as the linear dimension is reduced then it is realized that when the vacuole of a spherical cell of radius $30\ \mu$ shrinks to half its volume in say 5 minutes the passage of water is only 1 ml per $10,000\ \text{cm}^2$ per minute although the thickness of the layer between vacuole and external solution is of the order of $1\ \mu$ in thickness. Under other circumstances, this layer might be said to be relatively impermeable to water. It seems probable that much of the resistance resides not in the cellulose wall or cytoplasm but in the tonoplast which separates the latter from the vacuole.

G. E. BRIGGS

P

PARAMAGNETISM

A substance is said to be paramagnetic if it is attracted to a magnetic field in contradistinction to a diamagnetic substance which is repelled by a magnetic field. Excluding FERROMAGNETISM, which is an extremely strong type of paramagnetism, paramagnetic substances can be roughly divided into 3 categories—strong, weak, and very weak paramagnets.

Strong paramagnetism is exhibited by elements of the iron, platinum, palladium, rare earth, and uranium groups and (generally) compounds formed from elements of these groups. For such elements the paramagnetism arises from unpaired electron spins associated with the non-valence electrons of the atoms, each unpaired electron behaving much as a particle having an intrinsic spin angular momentum and associated magnetic moment which is orbiting about the central nucleus. For compounds in which the paramagnetic ions are dispersed sufficiently so that there are only negligible interactions between the paramagnetic electrons, the magnetic susceptibility, which is the ratio of the induced magnetization of the sample, M , to the externally applied magnetic field strength, H , is given by $\chi = C(T - \theta)$ for small values of H/T , T being the absolute temperature of the sample. In this equation, called the Curie-Weiss Law after P. Curie (1905) and P. Weiss (1907), who did much of the early work on paramagnetism, C is called the Curie constant and was shown by Langevin (1905), using classical arguments, to be equal to $N\mu^2/3k$, N being the number of paramagnetic ions per unit volume, μ the magnetic moment of each ion, and k Boltzmann's constant. θ is called the Curie or Curie-Weiss temperature and may be either positive, negative, or zero depending on the particular compound being studied. If $\theta = 0$ the Curie-Weiss Law reduces to the Curie Law $\chi = C/T$. For $T \lesssim \theta$, interactions among the ions can no longer be ignored and the Curie-Weiss Law is no longer valid. Later quantum mechanical calculations have shown that in the limit where $\mu H/kT \ll 1$, the Curie-Weiss Law is still valid providing that in the Curie constant C , μ is replaced by

$$\mu_{\text{eff}} = g\mu_B \sqrt{J(J+1)}$$

where g is the Lande g factor, having a value of about 2, μ_B is the Bohr magneton $|e|\hbar/2mc$, and J

is the total angular momentum quantum number of the ion in question. For some ions, notably those of the iron group, the contribution of the orbital angular momentum is partially or totally quenched by the crystalline electric field, in which case μ_{eff} is better given by the equation $\mu_{\text{eff}} = g\mu_B \sqrt{S(S+1)}$ where S is the spin angular momentum quantum number of the paramagnetic ion. A more complete derivation of χ for arbitrary values of H/T and for negligible interactions gives for the classical derivation (Langevin, 1905)

$$\chi = \frac{N\mu}{H} L\left(\frac{\mu H}{kT}\right), \quad L\left(\frac{\mu H}{kT}\right)$$

being the Langevin Function of the argument $\mu H/kT$. A quantum mechanical derivation gives

$$\chi = \frac{Ng\mu_B J}{H} B_J\left(\frac{g\mu_B JH}{kT}\right), \quad B_J$$

being the Brillouin Function. The Langevin and Brillouin Functions are given by the formulas

$$L(x) = \coth x - \frac{1}{x} \text{ and } B_J(x)$$

$$\frac{2J+1}{2J} \coth \left[\frac{(2J+1)}{2J} x \right] - \frac{1}{2J} \coth \left[\frac{x}{2J} \right]$$

A plot of $L(\mu H/kT)$ vs $\mu H/kT$ is shown in Fig. 1.

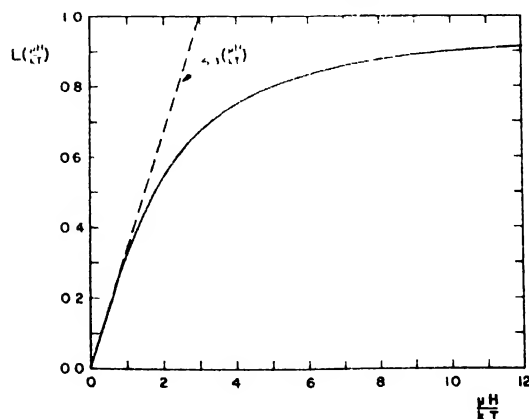


FIG. 1

It is noticed that for small values of $\mu H/kT$ the Langevin Function reduces to $1/3(\mu H/kT)$ which gives the Curie Law, while for large values of $\mu H/kT$ the Langevin Function approaches an asymptotic limit, which results in the saturation of the magnetization of the sample. A plot of the Brillouin Function for a given ion has the same qualitative shape as that shown for the Langevin Function. At room temperature the strong paramagnetism of the electrons ranges from 10^{-3} to 10^2 cgs units.

Weak paramagnetism is a strictly quantum mechanical phenomenon exhibited primarily by metals and is associated with the conduction electrons near the top of the Fermi distribution. Only those electrons having an energy within approximately $2\mu H$ of the Fermi energy are able to reorient or align themselves in a magnetic field and change their energy state, since all energy states below the Fermi energy are filled. Thus only the fraction of the total number of electrons kT/kT_F , where kT_F is the Fermi energy and T is the absolute temperature of the sample, can contribute to the susceptibility. Ignoring interactions among the electrons the susceptibility is then given by

$$\chi = \frac{C}{T} \cdot \frac{kT}{kT_F} = \frac{C}{T_F}$$

which is independent of temperature. A more rigorous quantum mechanical derivation of the susceptibility for this case gives the result that

$$\chi = \frac{3 N \mu_B^2}{2 k T_F} - \frac{1}{3} \left(\frac{3 N \mu_B^2}{2 k T_F} \right) = \frac{N \mu_B^2}{k T_F}$$

where the first term is the paramagnetic contribution associated with the electrons assumed to be stationary (Pauli, 1927), while the second term is the diamagnetic contribution due to the spatial motion of the electrons (Landau, 1930). At room temperature the susceptibility of the conduction electrons in a metal ranges from about 10^{-4} to 10 cgs units.

The very weak paramagnetism associated with some nuclei is due to their intrinsic spin angular momentum and associated magnetic moment. The nuclear susceptibility is described by the Brillouin equation as given above with J replaced by the nuclear spin quantum number I . At room temperature the nuclear susceptibility is approximately 10^{-11} to 10^{-10} cgs units and so is generally swamped by the paramagnetism of the electrons if such is present.

Electron paramagnetism is often studied experimentally using the Faraday or Gouy methods or the indirect method of electron paramagnetic resonance. The only direct measurement of nuclear paramagnetism was performed by Lasarew and Shubnikov (1937) for protons in solid hydrogen, although the sensitive techniques of nuclear magnetic resonance have now been used to study most known paramagnetic nuclei.

RONALD A. LAING

References

- Feynman, R. P., Leighton, R. B., and Sands, M., "The Feynman Lectures on Physics," Vol. 2, Reading, Mass., Addison-Wesley, 1964.
 Stoner, E. C., "Magnetism," Fourth edition, London, Methuen, 1948.
 Bates, L. F., "Modern Magnetism," Third edition, London, Cambridge, 1951.
 Kittel, C., "Introduction to Solid State Physics," Second edition, New York, John Wiley & Sons, Inc., 1956.
 Van Vleck, J. H., "Theory of Electric and Magnetic Susceptibilities," London, Oxford, 1932.

Cross-reference: MAGNETISM.

PARITY. See CONSERVATION LAWS AND SYMMETRY.

PERIODIC LAW AND PERIODIC TABLE

When the chemical elements are compared in order of increasing atomic number, many of their physical and chemical properties are observed to vary periodically rather than randomly or steadily. This relationship, recognized empirically a century ago by de Chancourtois in France, Newlands in England, Lothar Meyer in Germany, and Mendeleev in Russia, is now known to be the logical and inevitable consequence of the fundamental periodicity of atomic structure. The familiar statement of the periodic law is this: "The properties of the chemical elements vary periodically with their atomic number." A more informative statement of this same law is: *The atomic structures of the chemical elements vary periodically with their atomic number; all physical and chemical properties that depend on atomic structure therefore tend also to vary periodically with atomic number.*

The periodicity of atomic structure (see ATOMIC PHYSICS) is described by quantum theory as developed through modern wave mechanics. Each successive electron, beginning with the first, that comes within the field of an atomic nucleus, occupies the most stable position available to it. The number of possible positions is limited by quantum restrictions which describe each position in terms of four quantum numbers, and by the Pauli exclusion principle that no two electrons within the same atom may have the same four quantum numbers. These electron positions, or energy levels, are grouped with respect to their average distance from the nucleus as "principal quantum levels or shells," designated by the "principal quantum number" $n = 1, 2, 3, 4, \dots$, successive integral values increasing in order of increasing average distance from the nucleus. The total capacity of each shell can be expressed as $2n^2$, being 2 for $n = 1$, 8 for $n = 2$, 18 for $n = 3$, and 32 for $n = 4$; no higher level actually contains more than 32 electrons.

These total capacities can easily be accounted for by the several quantum number restrictions and the Pauli principle. Within each principal

energy level are differently shaped regions called "orbitals," that can be occupied by electrons. The shape of each orbital is designated by the "orbital quantum number" l , which may only have integral values from 0 up to $n - 1$. The number of orbitals of each shape that can exist within a principal quantum level depends on the fact that an electron in an orbital is a charge in motion and therefore has magnetic properties which influence the orientation of the orbital in an external magnetic field. The possible orientations are designated by the "orbital magnetic quantum number" m_l , which may have values from 0 to plus or minus the orbital quantum number l . Thus when $n = 1$, l can only have the value 0, which means that only one orbital is possible, having orbital magnetic quantum number 0. When $n = 2$, l can have the values 0 and 1. For the value 0, one orbital is possible, but when $l = 1$, m_l can have values 0, $+1$, and -1 , corresponding to three orbitals. Four orbitals are therefore possible in the principal quantum level $n = 2$. When $n = 3$, the same kinds of four orbitals are possible, and in addition, l can equal 2. This gives five possible values for m_l : 0, $+1$, $+2$, -1 , and -2 , corresponding to five more orbitals for a total of 9. When $n = 4$, the same kinds of 9 orbitals are possible, and in addition l can equal 3. This gives seven possible values for m_l : 0, $+1$, $+2$, $+3$, -1 , -2 , and -3 , corresponding to 7 more orbitals for a total of 16. No principal quantum level uses more than 16 orbitals even though more are theoretically possible when $n = 5$ or more.

One additional property of an electron in an atom needs to be considered. This is its property as a magnet, irrespective of its orbital motion. This is designated by the "spin magnetic quantum number," which can have only the values $+\frac{1}{2}$ and $-\frac{1}{2}$. Since each orbital is uniquely specified by the first three quantum numbers, n , l , and m_l , the capacity of each orbital is thus limited to 2 electrons, and these only if, according to the Pauli principle, they are of opposed spins (differ in the fourth quantum number). The total capacity of each principal quantum level is therefore twice the number of orbitals within it, because this represents the total number of permissible combinations of four quantum numbers within that level. For example, the capacity of the $n = 4$ level is limited to 32 electrons by the fact that only 16 different combinations of the four quantum numbers are possible within that level; these electrons will occupy 16 orbitals.

The differently shaped orbitals having orbital quantum numbers $l = 0, 1, 2$, and 3 are commonly called s , p , d , and f orbitals. From the above discussion, it should be clear that within any principal quantum shell, there can be only one s orbital, three p orbitals, five d orbitals, and seven f orbitals. The p orbitals do not appear until $n = 2$, the d until $n = 3$, and the f until $n = 4$. Within any given principal quantum shell, the order of decreasing stability, and therefore the order of filling with electrons, is always s - p - d - f .

The periodicity of atomic structure arises from the recurrent filling of new outermost principal

quantum levels, but it is complicated by the fact that although the principal quantum levels represent very roughly the general order of magnitude of energy, there is considerable overlapping. This overlapping is such that the outermost shell of an isolated atom can never contain more than 8 electrons. In the building up of successively higher atomic numbers, electrons always find more stable positions, once a set of p orbitals in a given principal quantum level is filled, in the s orbital of the next higher principal quantum level rather than the d orbitals of the same principal quantum level. When this s orbital is filled, electrons then go into the underlying d orbitals until these are filled, before continuing to fill the outermost shell by entering its p orbitals. The building-up of the atoms of successive atomic numbers may be represented by the following sequence: $1s, 2s, 2p, 3s, 3p, 4s, 3d, 4p, 5s, 4d, 5p, 6s, 5d, 4f, 6p, 7s, 6d, 5f$. The periodicity of atomic structure thus consists of the recurrent filling of the outermost shell with from one to 8 electrons that corresponds to the steady increase in nuclear charge.

A period is considered to begin with the first electron in a new principal quantum shell and to end with the completion of the octet in this outermost shell, except, of course, for the very first period, in which the outermost shell is filled to capacity with only two electrons. From the order of orbital filling given above, it should be apparent that periods so defined cannot be alike in length. The first period, consisting of hydrogen and helium, has only two elements. The second period, beginning with lithium (3) and ending with neon (10), contains 8 elements, as does the third period, which begins with sodium (11) and ends with argon (18). The fourth period begins with potassium (19), but following calcium (20), the filling of the outermost (fourth) shell octet is interrupted by the filling of the d orbitals in the third shell. Thus 10 more elements enter this period before filling of the outermost shell is resumed, making the total number of elements in this period 18. In the fifth period, the first two outermost electrons are added in rubidium (37) and strontium (38), but then this outer shell filling is interrupted by the filling of penultimate shell d orbitals, which again adds 10 elements before filling of the outermost shell is resumed; this period also contains 18 elements. The sixth period begins as before with two electrons in the outermost shell, but then there is an interruption at lanthanum (57) to begin filling the $5d$ orbitals. Here, however, occurs an additional interruption, in which 14 elements are formed through filling of the $4f$ orbitals, before the remaining $5d$ orbitals can be filled, and in turn before the outermost shell receives any more electrons. Consequently here it takes 32 elements to bring the outermost shell to 8 electrons and thus end the period. The seventh period is similar but incomplete. In principle it would end with element 118, but artificial element 103 is the highest in atomic number known at the time of writing.

These elements which represent interruptions in the filling of the outermost s and p orbital octet

are called "transition elements" (where d orbitals are being filled), and "inner transition elements" (where f orbitals are being filled). This is to distinguish them from the other, "major group," elements in which underlying d or f orbitals are either completely empty or completely filled.

Physical properties of the elements that depend only on the electronic structure of the individual atom, such as the ionization potential and atomic radius, vary periodically with atomic number simply because of the recurrent filling of the outermost shell. Increasing the atomic number, by increasing the nuclear charge while adding to the outermost shell electrons increases the attractive

interaction between nucleus and outermost electrons more than it increases the repulsive forces among the electrons, with the result that the electronic cloud tends to be held closer (smaller radius) and more tightly (higher ionization energy), the more outermost shell electrons there are. For example, carbon (6) with half-filled octet has radius and ionization energy intermediate between the larger lithium (3) atoms with low ionization energy and the smaller fluorine (9) atoms with high ionization energy. Similar but smaller effects are observable for the addition of d or f electrons to underlying shells.

The bonding properties of elements also depend

MAJOR GROUPS

No.	M 1	M 2	M 2	M 3	M 4	M 5	M 6	M 7	M 8	
1					H 1					He 2
2	Li 3	Be 4		B 5	C 6	N 7	O 8	F 9	Ne 10	
3	Na 11	Mg 12		Al 13	Si 14	P 15	S 16	Cl 17	Ar 18	
4	K 19	Ca		Zn	Ga 31	Ge 32	As 33	Se 34	Br 35	Kr 36
5	Rb 37	Sr		Cd	In 49	Sn 50	Sb 51	Te 52	I 53	Xe 54
6	Cs 55	Ba		Hg	Tl 81	Pb 82	Bi 83	Po 84	At 85	Rn 86
7	Fr 87	Ra								

TRANSITION

	T 3	T 4	T 5	T 6	T 7	T 8	T 9	T 10	T 11	
4	Sc 21	Ti 22	V 23	Cr 24	Mn 25	Fe 26	Co 27	Ni 28	Cu 29	
5	Y 39	Zr 40	Nb 41	Mo 42	Tc 43	Ru 44	Rh 45	Pd 46	Ag 47	
6	La 57	Lu	Hf 72	Ta 73	W 74	Re 75	Os 76	Ir 77	Pt 78	Au 79
7	Ac 89	Lr								

INNER TRANSITION

6	Ce 58	Pr 59	Nd 60	Pm 61	Sm 62	Eu 63	Gd 64	Tb 65	Dy 66	Ho 67	Er 68	Tm 69	Yb 70
7	Th 90	Pa 91	U 92	Np 93	Pu 94	Am 95	Cm 96	Bk 97	Cf 98	Es 99	Fm 100	Md 101	No 102

COMPLETE LONG FORM

1																	H							He								
2	Li	Be																	B	C	N	O	F	Ne								
3	Na	Mg																	Al	Si	P	S	Cl	Ar								
4	K	Ca	Sr													Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr		
5	Rb	Sr	Y													Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe		
6	Cs	Ba	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
7	Fr	Ra	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr															

FIG. 1. Periodic chart of the chemical elements (reprinted from *J. Chem. Educ.*, 188 (1964); copyright 1964 by Division of Chemical Education, American Chemical Society, and reprinted by permission of the copyright owner).

on the electronic structure of the individual atom and therefore likewise vary periodically. For example, each period (except the first) begins with an alkali metal, lithium (3), sodium (11), potassium (19), rubidium (37), cesium (55), and francium (87), all of which show similar metallic bonding, crystallizing in a body-centered cubic lattice. Each has but one outermost electron per atom and can therefore form but one covalent bond. Each is very low in electronegativity and thus tends to become highly positive when bonded to another element. Crossing each period the elements become less metallic, higher in electronegativity, and able to form a greater number of bonds until limited by the number of outer shell vacancies rather than the number of electrons. The halogens, fluorine (9), chlorine (17), bromine (35), iodine (53), and astatine (85), each of which is next to the end of its period, are all nonmetals, highest of their respective periods in electronegativity and thus tending to become highly negative when bonded to other elements. Having seven, outermost electrons, each has but one vacancy permitting but one covalent bond.

Physical properties of the elements that are of greatest interest are usually properties that depend indirectly on the atomic structure but directly on the nature of the aggregate of atoms which results from the atomic structure. Such properties are melting point, density, and volatility. They may tend to vary periodically, but the periodicity is not necessarily consistent or even evident, because of abrupt differences in the type of polyatomic aggregate. For example, the identical

atoms of carbon may form graphite or diamond, depending on the kind of bonding, with entirely different properties resulting. Nitrogen follows carbon in atomic number, but because it forms N_2 molecules instead of giant three-dimensional structures like diamond or graphite, it is a gas with physical properties strikingly different from those of either form of carbon.

Among the most useful applications of the periodic law is to an understanding of the differences among compounds, whose properties also vary in a periodic manner. For example; oxides of elements at the beginnings of periods tend to be very stable, high-melting, nonvolatile solids of strongly basic character and practically no oxidizing power. Oxide properties change progressively across each period until toward the end, the oxides tend to be unstable, low-melting, volatile compounds, acidic in nature and of high oxidizing power. Such periodicity is recognizable throughout a very large part of chemistry.

In order to erect a framework upon which the myriad facts that accord with the Periodic Law can be organized, the chemical elements can be arranged in an orderly manner called a "periodic table." Any such arrangement¹ can be satisfactory if it organizes the elements in some order of increasing atomic number, showing the separate periods and at the same time grouping elements of greatest similarity together.

In Fig. 1 is shown a modern version of the periodic table², together with the "long form." Figure 2 shows the "long form" as currently most widely used. From left to right across the table are

																INERT GASES																										
IA	IIA	IIIB	IVB	VB	VIB	VII	VIII	IB	IIB	IIIA	IVA	VA	VIA	VIIA	VIIIA	1	2																									
1 H 1.00797																1 H 1.00797	2 He 4.0026																									
3 Li 6.939	4 Be 9.0122															9 F 18.9984	10 Ne 20.183																									
11 Na 22.9898	12 Mg 24.312															17 Cl 35.453	18 Ar 39.948																									
19 K 39.102	20 Ca 40.08	21 Sc 44.956	22 Ti 47.90	23 V 50.942	24 Cr 51.996	25 Mn 54.9380	26 Fe 55.847	27 Co 58.9332	28 Ni 58.71	29 Cu 63.54	30 Zn 65.37	31 Ga 69.72	32 Ge 72.59	33 As 74.9216	34 Se 78.96	35 Br 79.904	36 Kr 83.80																									
37 Rb 85.47	38 Sr 87.62	39 Y 88.906	40 Zr 91.22	41 Nb 92.906	42 Mo 95.94	43 Tc (98)	44 Ru 101.07	45 Rh 102.905	46 Pd 106.4	47 Ag 107.870	48 Cd 112.40	49 In 114.82	50 Sn 118.69	51 Sb 121.75	52 Te 127.60	53 I 126.9044	54 Xe 131.30																									
55 Cs 132.905	56 Ba 137.34	57 La 138.91	72 Hf 178.49	73 Ta 180.948	74 W 183.85	75 Re 186.2	76 Os 190.2	77 Ir 192.2	78 Pt 195.08	79 Au 196.967	80 Hg 200.59	81 Tl 204.37	82 Pb 207.19	83 Bi 208.980	84 Po (210)	85 At (210)	86 Rn (222)																									
87 Fr (223)	88 Ra (226)	89 Ac (227)	Lanthanum Series												58 Ce 140.12	59 Pr 140.907	60 Nd 144.24	61 Pm (147)	62 Sm 150.35	63 Eu 151.96	64 Gd 157.25	65 Tb 158.924	66 Dy 162.50	67 Ho 164.930	68 Er 167.26	69 Tm 168.934	70 Yb 173.04	71 Lu 174.97														
																	Actinium Series												90 Th 232.038	91 Pa (231)	92 U 238.03	93 Np (237)	94 Pu (242)	95 Am (243)	96 Cm (247)	97 Bk (247)	98 Cf (248)	99 Es (254)	100 Fm (253)	101 Md (258)	102 No (259)	103 Lw (261)

Fig. 2. Periodic chart of the elements (reproduced with permission of copyright owner, Fisher Scientific Company).

the "periods," representing the elements in order of successively increasing atomic number and therefore progressive change in atomic structure. Across the period, the properties of the elements and their compounds tend to change from one extreme to the opposite. From top to bottom the periods are placed so that the most similar elements are grouped together. A distinction (shown by a physical separation) is made between "major groups" and "transition groups" because of the bonding dissimilarities originating in the availability, in the transition elements, of underlying *d* orbitals that are not available in the major group elements. In older tables, major groups are usually, although inconsistently, designated as "A" and transition groups as "B." In Fig. 1, the designations are consistently "M" (major) and "T" (transition). But whether in the major group or the transition group, the elements are arranged vertically on the basis of similarity in electronic configuration, which is then reflected in similarities in the physical and chemical properties of the elements and their compounds.

No amount of organization or correlation will ever alter the fact that each chemical element is an individual and unique, nor will the properties of any element be changed one iota by placing that element in any special position in a periodic table. Nevertheless, there is enough consistency to the structure and behavior of atoms to make any reasonable form of the periodic table an extremely useful framework upon which to organize and correlate an enormous quantity of chemical information³. The periodic law is truly one of the great generalizations of science.

R. T. SANDERSON

References

1. Mazurs, E., "Types of Graphic Representation of the Periodic System of Chemical Elements," published by the author, 6 S. Madison Ave. La Grange, Ill. 1957.
2. Sanderson, R. T., *J. Chem. Educ.*, **41**, 187 (1964).
3. Sanderson, R. T., "Chemical Periodicity," New York, Reinhold Publishing Corp., 1960.

Cross-references: ATOMIC PHYSICS; ELECTRON; ELEMENTS, CHEMICAL; ISOTOPES; QUANTUM THEORY; TRANSURANIUM ELEMENTS.

PHASE RULE

The phase rule is a general equation $F = n - r - 2$, stating the conditions of thermodynamic equilibrium in a system of chemical reactants. The number of degrees of freedom or variance (F) allowed in a given heterogeneous system may be examined by analysis or observation and plotted on a graph by proper choice of the components (n), the phases (r), and the independently variable factors of temperature and pressure.

Josiah Willard Gibbs propounded the rule about 1877, and H. W. B. Roozeboom about 1890 began pioneering in specific cases. This rule

has been important in the development of metallurgy, the exploitation of salt deposits (e.g. Stassfurt, Germany and Searles Lake, Calif.), and the study of ceramic and mineralogic processes.

A phase is a homogeneous, physically distinct and mechanically separable portion of a system. H₂O has the three phases: water, vapor and ice. Each crystal form present is a phase. The relative amounts of each phase do not affect the EQUILIBRIUM.

The one-component system, water-vapor-solid, is unary. The components of a system are the smallest number of independently variable constituents by means of which the composition of each phase taking part in the equilibrium can be expressed in the form of a chemical equation.

With the components fixed in a system, variance—the degrees of freedom (F)—depends on the number of phases present. If water vapor alone is present, the system is bivariant since both temperature and pressure can vary within limits without affecting the number of phases; but if a second phase, liquid water, is present, the system is univariant and if either the temperature or pressure of the system in equilibrium is set, the other is automatically fixed as long as a second phase is present. A third phase, ice, makes the system invariant (the triple point), and any change in temperature or pressure, if maintained, results in the disappearance of one phase. Addition of another component forms a binary system, one degree of freedom is added, and the system, is univariant until a fourth phase appears and the system becomes invariant (the quadruple point).

Schematic phase equilibrium diagrams outline experimental observations of physical and chemical changes in the system as the conditions of temperature, pressure and composition are varied. For unary systems, the diagram has two dimensions, for binary systems, three, and for ternary systems, four, etc. Binary systems are easily plotted with pressure or temperature constant, while ternary systems may be treated similarly as condensed systems with both constant if the vapor pressure is less than atmospheric. This added restriction reduces the variance by one, and at constant temperature, composition relationships may be plotted on a triangular diagram. Quarternary or quinary systems and ternary systems above atmospheric pressure may be treated by projections of surfaces of thermodynamic stability, but more complex systems require a mathematical approach.

The simple one-component system, water, plotted with rectangular coordinates, i.e., pressure and temperature, shows a variety of concepts which may be extended to more complex systems (see Fig. 1 in STATES OF MATTER). Each area in the diagram is a bivariant, one-phase state. Each curve separating the areas is a univariant, two-phase state showing the conditions under which a transition of phase occurs. The fusion curve for the equilibria between the solid phase and the liquid phase, the sublimation curve for solid

and vapor, and the vaporization curve for liquid and vapor meet at the triple point. The three distinct phases differ in all properties except chemical potential. The end of the vaporization curve is a singular point, the critical point where liquid and vapor become identical, a restriction which reduces the variance by one so that F becomes zero.

In a binary (or higher) system, a liquidus is a curve representing the composition of the equilibrium liquid phase and a solidus represents the composition of the solid phase. The conjugate vapor phase is represented by a vaporous with tie-lines or conodes to the liquidus or solidus points in equilibrium. A minimum point for the existence of a liquid is the eutectic, sometimes (in an aqueous system) called the cryohydric point. In a ternary system the eutectics of three binary systems initiate curves leading to a ternary eutectic. If two liquids form a miscibility gap in the system, multiple quadruple points are possible. At a peritectic, a phase transition occurs at other than a minimum, i.e., one solid melts to another solid and a liquid. The solid may be a compound or a solid solution (mixed crystals) in which the composition of the solid varies with relative proportion of components and is shown by the solidus. A congruent melting point is a maximum in its curve, i.e., the solid melts to a liquid of the same composition. Where two phases become identical in composition, an indifferent or critical point exists. Where two conjugate phases become identical, the point may also be called a consolute point.

JOHN H. WILLS

References

- Ricci, John E., "The Phase Rule and Heterogeneous Equilibrium," New York, D. Van Nostrand Co., 1951.
- Levin, E. M., McMurdie, H. F., and Hall, F. P., "Phase Diagrams for Ceramists," Columbus, Ohio, The American Ceramic Society, 1956.
- Masing, G., "Ternary Systems," New York, Reinhold Publishing Corp., 1944.
- Wetmore, F. E. W., and LeRoy, D. J., "Principles of Phase Equilibria," New York, McGraw-Hill Book Co., 1951.

Cross-references: EQUILIBRIUM STATES OF MATTER, THERMODYNAMICS.

PHONONS

Many of the thermal and vibrational properties of solids can be explained by considering the material to be a volume made up of a gas of particles called phonons. This particle description is a method of taking into account the actual motion of the atoms and molecules in the solid. Since each atom possesses energy due to its thermal environment, and since there are forces between the atoms which keep the solid together, each atom tends to oscillate about its equilibrium

position. The formal mathematical development, obtained through solving the equations of motion of the array of individual atoms and molecules, indicates that the thermal energy of the solid is contained in certain combinations of particle vibrations which are equivalent to standing elastic waves in the sample and are called normal modes. Each normal mode contains a number of discrete quanta of energy $E = \hbar\omega$, where ω is the frequency of the mode (or wave) and \hbar is Planck's constant divided by 2π . Each of these quanta is called a phonon (in analogy with the light quanta or photon whose energy-frequency relationship is identical). The remaining discussion can be best understood by considering the phonons only as particles, each having an energy $E = \hbar\omega$, a momentum q , and a velocity $v = \partial\omega/\partial q \sim \omega/q$ (see QUANTUM THEORY).

Analogous to the energy levels of electrons in a solid, phonons can have only certain allowed energies. A typical phonon spectrum, which relates the energy of the phonons to their momentum (or wave vector), is shown in Fig. 1. In this

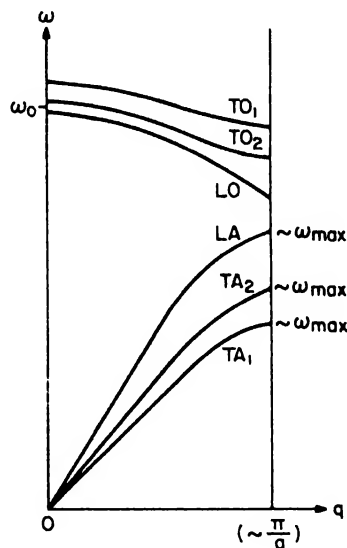


FIG. 1. Phonon spectrum.

Figure, the phonon energy is $\hbar\omega$ and the momentum (or wave vector) is q . There are $3N$ discrete values of q associated with a lattice containing N particles, and the maximum value of q is of the order of (π/a) ($\sim 10^8 \text{ cm}^{-1}$) where a^3 is the volume of a unit cell. (A unit cell is the approximate volume taken up by each molecule of the crystal.) Typical values of the points marked ω_{\max} are 10^{12} to 10^{13} radians/sec.

The two types of phonon branches shown in Fig. 1 (A and O), are the acoustical branch ($\omega = 0$ at $q = 0$) and the optical branch. The motion of the atoms which is represented by the acoustical branch is similar to that which is obtained when a sound wave is propagated in the crystal. The optical branch represents motion in

which the center of mass of the different atoms in a unit cell (or a molecule) remains fixed, but the atoms in the molecule move relative to each other. This type of motion is often excited by an electric field or a light wave, hence the name optical phonon.

Each branch is further divided into longitudinal (L) and transverse (T) modes, again in analogy with propagation of sound waves. In the longitudinal branch, the atoms vibrate in the same direction that the phonon moves, while in the transverse branches, the atoms vibrate perpendicular to the phonon motion. It should be noted that there is a forbidden energy gap between LA and LO which indicates that phonons having these frequencies do not exist in the material.

Several equations which represent the energy content of the phonon modes complete the description. The total energy of each mode q is quantized according to

$$E_q = n_q \hbar \omega_q \quad (1)$$

where n_q is the number of quanta excited at temperature T with frequency ω_q and is given by

$$n_q = [\exp(\hbar \omega_q / kT) - 1]^{-1} \quad (2)$$

where k is Boltzmann's constant. By combining Eq. (1) and (2), we note that at any temperature T , most of the phonons have an energy of about kT , as long as $kT/\hbar < (\omega_q)_{\max}$. The value of $T = (\hbar \omega_q)_{\max} / k$ is approximately θ_D , the Debye temperature.

For completeness, a density term $\rho_j(\omega)$ should be defined. This represents the fraction of phonons of the j th branch which have frequencies between ω and $\omega + d\omega$. However, the form of the distribution is complicated, depending on the symmetry of the lattice and the strength of the restoring forces between atoms. For low-frequency acoustical phonons, in a three-dimensional lattice,

$$\rho_j(\omega) \propto \omega^2 d\omega \quad (3)$$

In any process in which phonons interact with any other particles (or excitations) in the solid, for example electrons, photons, or neutrons, the normal conservation laws must be satisfied. That is, total energy and total momentum must be unchanged. Thus if a phonon is destroyed in an interaction with an electron, the electron must gain an amount of energy $\hbar \omega$ and momentum q which the phonon originally had.

Phonon Interactions and Transport. Many of the thermal properties of solids which are due to the lattice, as distinguished from the electrons, can be obtained using the energy of Eq. (1) in the normal thermodynamic relations. For example, the specific heat C_q due to all the phonons with momentum q is given by

$$C_q = \frac{\partial E_q}{\partial T} = k \left(\frac{\hbar \omega}{kT} \right)^2 \frac{e^{\hbar \omega / kT}}{(e^{\hbar \omega / kT} - 1)^2} \quad (4)$$

To obtain the total specific heat, we simply add up the contributions from each mode q . This

sum is most easily obtained if we know the distribution function $\rho_j(\omega)$ and the value of $(\omega_q)_{\max}$. For example, the Debye theory uses approximate expressions for these quantities and obtains

$$C = 9Nk(T/\theta_D)^3 \int_0^{\theta_D/T} \frac{x^4 e^x dx}{(e^x - 1)^2} \quad (5)$$

where N is the number of atoms per unit volume.

To obtain the heat conductivity in solids, the transport of energy by means of the phonons must be considered. Phonons can be very efficient heat carriers, as indicated by the fact that the maximum thermal conductivity due to electrons in copper is about 50 watts/cm deg as compared to about 200 watts/cm deg due to phonons in pure single crystal sapphire, in which all the heat is carried by phonons.

The heat conducted by the phonons of momentum q is given approximately by

$$\kappa_q \approx \frac{1}{3} C_q v_q l_q \quad (6)$$

The total thermal conductivity is obtained by summing the values of κ_q over all the possible values of q . The mean free path l_q is a measure of how far a phonon travels before it interacts, or is scattered. Phonons can be scattered by impurities and imperfections, by electrons, by other phonons and by the walls of the solid. The mean free path for any of these processes can be determined, to some extent, by examining the temperature dependence of the thermal conductivity.

Another method of investigating the scattering of phonons is through the investigation of the attenuation of very high-frequency sound waves in solids. It is possible to generate "ultrasonic" sound waves with frequencies in excess of 10^9 cps. These correspond to the lowest values of ω shown in Fig. 1 and would be, for example, the majority of the phonons if the sample were held at 0.5 K. In the near future these ultrasonic frequencies will be increased, and more of the phonon spectrum will be investigated using this tool.

The Determination of the Phonon Spectra. The actual values for the ω -vs- q curves which make up the phonon spectrum are obtained in a variety of ways. The value ω_0 of the $q = 0$ optical phonon is related to the infrared Reststrahl frequency. The low q values for the acoustic branches can be obtained by measuring the velocity of transverse and longitudinal sound waves in solids. The values of ω near the zone boundaries ($q \approx \pi/a$) can often be obtained from infrared absorption experiments. However, the most important tool for elucidating the details of the phonon spectrum has been slow neutron scattering. By examining the intensity and position of a scattered neutron beam, a large portion of the ω -vs- q curves can be mapped. Once the ω -vs- q curves have been obtained for a solid, it is possible by mathematical analysis to examine the nature and the magnitude of the forces between the atoms of that particular material. This has been done for a number of materials.

The phonon is of importance to a great many phenomena; electron mobility, optical absorption, electron spin resonance, electron tunnelling, and superconductivity are a few. At the same time, the phonon spectrum represents a detailed picture of the forces which hold solids together. Thus it is clear why the phonon has been and will continue to be of fundamental importance in solid-state physics.

M. G. HOLLAND

References

- Ziman, J., "Electrons and Phonons," Oxford, The Clarendon Press, 1960.
 Kittel, C., "Introduction to Solid State Physics," New York, John Wiley & Sons, 1956.
 Klemens, P. G., "Thermal Conductivity and Lattice Vibrational Modes," in Seitz, F., and Turnbull, D., Eds., "Solid State Physics," Vol. 7, p. 1, New York, Academic Press, 1958.
 Shull, C. G., and Wollan, E. O., "Applications of Neutron Diffraction to Solid State Problems," in Seitz, F., and Turnbull, D., Eds., "Solid State Physics," Vol. 2, p. 137, New York, Academic Press, 1956.

Cross-references. CRYSTALLOGRAPHY, ELECTRON, HEAT CAPACITY, HEAT TRANSFER, NEUTRON, PHOTON, QUANTUM THEORY, ULTRASONICS.

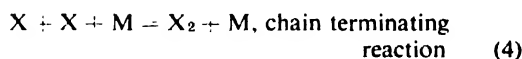
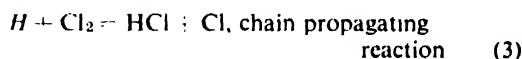
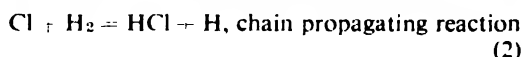
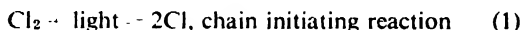
PHOTOCHEMISTRY*

Photochemistry deals in its broadest sense with the phenomena produced by energy-rich photoexcited states.¹ These states, when produced by the absorption of visible or ultraviolet light, are electronically excited states resulting from the transfer of an electron to a higher energy level. The time required for this transfer is so short that the positions of the nuclei of the atoms involved remain unchanged. This primary act is followed the movement of the nuclei to new equilibrium positions and by a great many different kinds of processes such as fluorescence, phosphorescence, degradation of the absorbed energy to heat, the intra- or intermolecular transfer of the electronic energy, or of electrons, protons or hydrogen atoms, or the breaking apart of the energy rich excited species, such as Cl_2 into Cl atoms, O_3 into O_2 and O, the production of a wide variety of ions, radicals, new molecules or chemical reactions. Such reactions are the natural photosynthetic process, photography, the electron transfer reaction in which gaseous hydrogen and oxygen are produced from water by light absorbed by cerium ions in water, photochromism as in the ortho-nitrotoluenes, *cis* to *trans* and *trans* to *cis* isomerization as in the case of the thioindigos, anils and azo compounds, shifting of

the positions of double bonds as occurs when ergosterol is converted into Vitamin D and addition reactions to multiple bonds.

The unit of light energy most useful in photochemistry is the photon, $\epsilon = hc/\lambda$, where h is Planck's constant ($6.5 \cdot 10^{-27}$ erg sec), c is the velocity of light (3×10^{10} cm/sec) and λ is the wavelength of the light, e.g., 6700 Å or $6700 \cdot 10^{-8}$ cm for the most intense light reaching the earth from the sun. Another photochemical unit of light energy is the einstein which is the energy of $6 \cdot 10^{23}$ or one mole, N , of light quanta. Thus one einstein of red light is $N\epsilon = Nhc/\lambda = 6 \cdot 10^{23} \times 6.5 \cdot 10^{-27} \times 3 \times 10^{10}/6700 \times 10^{-8} = 17.5 \cdot 10^{11}$ ergs or $17.5 \cdot 10^{11}/4.186 \cdot 10^7 = 42\,000$ calories or $42\,000/23\,024 = 1.8$ eV. This amount of energy is greater than the energy of activation required to initiate most thermal reactions.

In a narrower sense, photochemistry deals only with the chemical reactions brought about by the absorbed light. This includes studies of the kinetics and mechanisms of reactions such as the mechanism of the reaction between H_2 and Cl_2 to produce HCl. In this case only the Cl_2 molecules absorb visible light and the reaction proceeds mainly as follows:

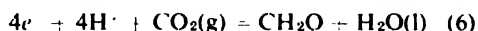


where X is either H or Cl. The net quantum yield for this chain reaction has been found under favorable conditions to be over one million moles of HCl produced per mole of visible light quanta absorbed by the chlorine molecules.

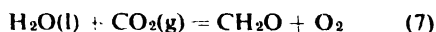
The natural photosynthetic process converts solar energy into energy available in storage for man's use. One part of the process is the oxidation of water to molecular oxygen by the over-all half reaction:



The other part of the process is the reduction of carbon dioxide to carbohydrate by the over-all half reaction:



The sum of reactions (5) and (6) is the charge and energy transfer process:



which is brought about by sunlight absorbed neither by the water nor the carbon dioxide but by the green substance, chlorophyll, in the proper watery environment.

The production of heat and light by the burning of wood in oxygen is the reverse of reaction (7).

*This is publication No. 90 of the M.I.T. Solar Energy Conversion Project which is supported by the Godfrey Lowell Cabot Fund of the Massachusetts Institute of Technology.

It thus becomes apparent that the energy produced in this way came originally from the sunlight which produced reaction (7) and thereby converted solar energy into energy available in storage for man's use.

The path of oxygen in the natural photosynthetic process has been followed by the use of the oxygen isotope of mass 18 and the path of carbon by the use of the radioactive isotope of mass 14. Very little is known, however, about the way the light absorbed by the chlorophyll brings about the reaction. Studies designed for this purpose would require the identification of all the light-absorbing species, including the transient intermediates, and their absorption spectra and absorbances.

One key to the elucidation of photochemical reactions is the determination of net quantum yields based on the amount of the measured reaction produced by light of nearly uniform intensity absorbed by the species which thereby initiated the measured reaction in an adequately mixed system under nearly equilibrium conditions. Especially helpful is the study of the influence upon the net yields of such variables as light intensity, temperature and the absolute and relative concentrations of the components of the system.

Net quantum yields are usually calculated from gross quantum yields which are based on the amounts of the measured effect and light absorbed by the entire system and on the absorption spectra, absorbances and concentrations of the different light absorbing species. Net quantum yields are evaluated most easily when the light is monochromatic, i.e., when the light consists of a narrow band of wavelengths such as the light of 2537 Å isolated from a low-pressure mercury vapor lamp.² This light, however, needs to be sufficiently intense to produce an appropriate amount of reaction in a convenient length of time.

Another key to the elucidation of photochemical reactions is the employment of flash photolysis made possible by apparatus consisting of a flash source and auxiliary apparatus for studying light absorbing intermediates of lifetimes as short as a few microseconds.²

LAWRENCE J. HEIDT

References

1. Bowen, E. J., "The Chemical Aspects of Light," Second edition, Oxford, The Clarendon Press, 1946.
2. Heidt, L. J., and Landi, V. R., *J. Chem. Phys.*, **41** 176 (1964).

Cross-reference: RADIATION CHEMISTRY.

PHOTOCONDUCTIVITY

Photoconductivity is the increase in electrical conductivity which occurs in a nonmetallic solid when it is exposed to electromagnetic radiation. The conductivity increase is due to the additional

free carriers which are generated when photon energies are absorbed in electronic transitions. The rate at which free carriers are generated and the length of time they persist in conducting states (lifetime) determines the amount of conductivity change.

The absorption transitions and, therefore, the photoconductivity resulting from them are termed intrinsic or extrinsic according to the energy states involved. When the photon energy is at least equal to the forbidden energy gap ($hc/\lambda \geq E_g$), electrons may be excited from the filled band to the conduction band. This intrinsic absorption transition produces an electron-hole pair, an electron in the conduction band and a hole in the valence band. Free carriers may be produced at lower photon energies when there are impurities or other crystal defects which give rise to energy states in the forbidden gap. Such a state is at an energy $E_c - E$ below the conduction band and $E - E_v$ above the valence band. Depending on whether the state is or is not occupied by an electron, a photon of appropriate energy may produce a transition from the state to the conduction band ($hc/\lambda \geq E_c - E$) or from the valence band to the impurity state ($hc/\lambda \geq E - E_v$). The former transition produces a free electron and the latter a free hole.

The carriers so generated are "excess" carriers, i.e., in excess of the number of carriers present in the solid in thermal equilibrium. In equilibrium, in the absence of light, the rate at which free carriers are generated by thermal processes and the rate at which they return to nonconducting ground states (recombination rate) are equal. The density of free carriers remains constant in time, aside from statistical fluctuations. The production of free carriers by absorption of radiation amounts to an increase in the total generation rate of free carriers, unbalancing the equilibrium. The recombination rate is a function of free-carrier density, so it will also increase as the free-carrier density increases. Thus, under steady illumination, there will be a new equilibrium (steady-state) at a new carrier density ($n_0 + \Delta n$), at which the rate of generation of carriers due to absorption (G) is equaled by the increase in the recombination rate (ΔU) due to the increased carrier density (Δn). The steady-state value of Δn is equal to the product $G\tau$ since the lifetime (τ) is equal to $\Delta n/\Delta U$.

The rate of change of excess carrier concentration is described by the equation

$$\frac{d\Delta n}{dt} = G - \frac{\Delta n}{\tau} = G - \Delta U \quad (1)$$

In general τ is not independent of Δn so ΔU is not always simply proportional to Δn . The dependence of τ on Δn is a property of the recombination mechanism. There are a variety of possible recombination mechanisms. Recombination may occur by a direct recombination of a free electron and a free hole or indirectly through a localized energy level (recombination center) which captures a hole and then an electron or

vice-versa. In extrinsic photoconduction, the localized energy state from which the free carrier was excited is, in effect, an immobilized carrier of the opposite type. In this case the recombination kinetics are essentially the same as for direct recombination. Several recombination processes may operate simultaneously. However, in most cases one will have the highest recombination rate and will dominate.

Direct Recombination. In direct recombination the recombination rate is the same for electrons and holes since they recombine in a single process. The total rate is given by

$$U = C(n_0 + \Delta n)(p_0 + \Delta p) \quad (2)$$

where C is the capture probability, n_0 and p_0 are the equilibrium densities of electrons and holes, and Δn and Δp are the excess carrier densities. Δn is equal to Δp since electrons and holes are generated and recombine in pairs. The equilibrium recombination rate is Cn_0p_0 so the excess recombination rate is

$$\Delta U = C(n_0 + p_0 + \Delta n)\Delta n \quad (3)$$

and the lifetime of excess carriers is

$$\tau = [C(n_0 + p_0 + \Delta n)]^{-1} \quad (4)$$

For small signals τ is independent of Δn . However, when Δn is comparable with $(n_0 + p_0)$, this is no longer the case. In insulators (where $n_0 + p_0$ is very small) or for large signals, $\Delta n \gg n_0 + p_0$. This leads to the steady-state relation

$$\Delta n = G\tau = \sqrt{G/C} \quad (5)$$

Recombination Centers and Traps. Indirect recombination through localized centers requires separate equations for electrons and holes, since the recombination rates for electrons and holes are not necessarily equal except in steady state. There are four transitions involved in the kinetics of localized centers: (1) A filled center may capture a hole from the valence band; (2) the electron may be thermally excited to the conduction band; (3) an empty center may capture an electron from the conduction band, or (4) the trapped hole may be excited to the valence band. The rate of capture of electrons by localized centers is given by $C(1-f)n$ where C is the capture probability of the centers for electrons, N is the density of centers, f is the fraction of centers already occupied by electrons, and n is the electron density. The product $(CN)^{-1} = \tau_{n0}$ is the characteristic lifetime due to capture by that center when n is the minority carrier in strongly p -type material. The rate of thermal excitation from the centers is given by $CfNn_0 \exp(-E/kT)$, where N_0 is the effective density of states in the conduction band and E is the thermal ionization energy for the centers. With $N_0 \exp(-E/kT) = n'$, the total capture rate for electrons may be written

$$U_n = \frac{(1-f)n}{\tau_{n0}} - \frac{fn'}{\tau_{n0}} \quad (6)$$

The expression for holes, with equivalent definitions, is

$$U_p = \frac{fp}{\tau_{p0}} - \frac{(1-f)p'}{\tau_{p0}} \quad (7)$$

In thermal equilibrium, these net capture rates are each equal to zero. When n and p are increased (f remaining constant for the moment) the capture rates for electrons and holes increase by $[(1-f)\Delta n]/\tau_{n0}$ and $(f\Delta p)/\tau_{p0}$, respectively. If f is not too nearly equal to unity and $\tau_{p0} \gg \tau_{n0}$, then $(1-f)/\tau_{n0} \gg f/\tau_{p0}$. With Δn and Δp being generated in equal densities, the capture rate for excess holes does not balance that for excess electrons. As a consequence, f must increase; i.e., electrons are trapped. If N is large, then the density of trapped electrons, $N\Delta f$, may be large. On the other hand $(1-f)/\tau_{n0}$ may be comparable to f/τ_{p0} in which case the center may function as an efficient recombination center. If N is not too large, a small amount of trapping may suffice to equilibrate the two capture rates.

The discussion of the basic photoconductive effect above has tacitly assumed uniform material, neutral contacts, and uniformly absorbed photon flux. Nonuniformity of material, either resistivity gradients or p - n junctions, gives rise to observable voltages due to diffusion of photoexcited carriers. This photovoltaic effect is used in solar cells and exposure meters. Nonuniformly absorbed radiation, as will occur when strongly absorbed radiation is absorbed entirely in a thin surface layer on a bulk crystal, will produce a diffusion of photoexcited carriers into the bulk of the crystal. When mobilities are different or trapping is present, a voltage may be detected (the Dember voltage) or, in the presence of a magnetic field, the photoelectromagnetic effect. Even when a photoconductor is otherwise uniform in composition, its surfaces introduce a perturbing influence. For example, the surface may contain a set of very efficient recombination centers, so that lifetime near the surface is much smaller than in the bulk. This will cause a nonuniformity in excess carrier concentration which will not be constant in time (in general). This surface recombination may dominate lifetime in a thin sample, and in a thick sample it may cause a deviation from exponential decay even if the bulk lifetime is independent of Δn .

The phenomenon of photoconductivity is intimately related to semiconductor device phenomena since most devices, e.g., transistors, lasers, luminescent materials, operate by means of non-equilibrium densities of carriers which are subject to the same recombination and trapping kinetics as in photoconductivity.

JOSEPH F. WOODS

References

- Bube, R. H., "Photoconductivity of Solids," New York, John Wiley & Sons, Inc., 1960.
- Tauc, J., "Photo- and Thermoelectric Effects in Semiconductors," New York, Pergamon Press, 1962.

- Levinstein, H., Ed., "Proceedings of the 1961 International Conference on Photoconductivity," *J. Phys. Chem. Solids*, **22** (1961).
- Moss, T. S., "Photoconductivity in the Elements," London, Butterworths, 1952.
- Kruse, P. W., McGlauchlin, L. D., and McQuistan, R. B., "Elements of Infrared Technology," New York, John Wiley & Sons, 1962.
- Breckenridge, R. G., Ed., "Photoconductivity Conference," New York, John Wiley & Sons, 1956.
- Rose, A., "Concepts in Photoconductivity," New York, Interscience Publishers, 1963.

Cross-references: CONDUCTIVITY, ELECTRICAL; ENERGY LEVELS; PHOTOELECTRICITY; PHOTON; PHOTOVOLTAIC EFFECT; SEMICONDUCTORS.

PHOTOELASTICITY

Photoelasticity is a method of determining stresses or strains in transparent materials by means of polarized light. It is widely used in engineering laboratories for the analysis of stress distributions in models of structures and machine parts. In the majority of problems, the results of such model tests are directly transferable to the metallic prototypes, as long as the materials of the models and prototypes both remain *elastic* (hence the origin of the term).

The *photoelastic effect* can be explained by assuming that an initially plane-polarized light beam, in passing through the stressed plane specimen, breaks up at every point into two components (*birefringence*), corresponding to the directions of the two principal stresses at that point (see POLARIZED LIGHT). The relative retardation between these two components will be proportional to the difference of principal stresses at each point. If the emergent light is passed through a second plane-polarizing unit (the *analyzer*), the principal plane of which is usually put at right angle to the first one, the two relatively retarded components are reduced into one plane and give rise to interference. In general, a number of such interference bands will appear in the model simultaneously, along each of which the principal stress difference (or maximum shear) will be of a constant value, that can be determined by calibration. If white light is used for the source, these bands will appear in the subsequent colors of the spectrum and are thus called *isochromatics*. At points where the incident plane-polarized light happens to coincide with the direction of one of the principal stresses, the light will merely pass through as a plane wave and, subsequently, will be cut out by the crossed analyzer. All such points in the piece will thus be connected with a dark band for any particular inclination of the crossed polarizer and analyzer. These bands are the *isoclinics*, which define the orientation of the principal stresses. They can be eliminated by the use of quarter-wave-plates, which produce circularly polarized light, devoid of the above directional property.

The isochromatics and isoclinics furnish two of the three independent parameters needed to

define the state of stress at any point in a two-dimensional field. The third parameter may be obtained by a variety of supplementary means such as lateral extensometry, mechanical or electrical analogies, or by numerical integration.

The increased use of the photoelastic method was greatly aided by successive improvements in model materials, from glass and celluloids to the latest thermosetting resins, such as the epoxies.

An extension of the method to spatial problems (three-dimensional photoelasticity) provides the only known experimental means for the determination of the state of stress in the interior of solids. The method is also used in dynamic (transient stress) studies, and its scope is extended now into nonelastic fields (photoplasticity and photoviscoelasticity).

M. HETÉNYI

References

- Coker, E. G., and Filon, L. N. G., "A Treatise on Photo-Elasticity," Cambridge, Cambridge University Press, 1931.
- Hetényi, M., "The Fundamentals of Three-Dimensional Photoelasticity," *J. Appl. Mech.*, **5**, No. 4 (1938).
- Frocht, M. M., "Photoelasticity," Vol. I (1941), Vol. II (1948), New York, John Wiley & Sons.
- Hetényi, M., Ed., "Handbook of Experimental Stress Analysis," New York, John Wiley & Sons, 1950.
- Hetényi, M., "Photoelasticity and Photoplasticity," in Goodier, J. N., and Hoff, J., Eds., "Structural Mechanics," New York, Pergamon Press, 1960.

Cross-references: ELASTICITY, INTERFERENCE AND INTERFEROMETRY, POLARIZED LIGHT.

PHOTOELECTRICITY

For the purposes of this article, photoelectricity is defined as the emission of electrons resulting from the absorption of radiation in a material. This definition is purposely broad: by radiation, we understand electromagnetic radiation ranging from the shortest wavelength to the longest, from gamma-rays into infrared. The electrons can be released into vacuum or into a second material and the material itself can be either a solid, a liquid, or a gas.

Although there may have been some observations, which with hindsight could be called photoelectric observations, the real history of photoelectricity starts with the work of Hertz. It was in the course of his experiments on electrical resonance that he observed that the length of the spark which could be induced in an auxiliary circuit was greater if the spark gap was irradiated by the ultraviolet light generated by the spark of the primary circuit. He found also that the effect is most marked when the electrode, which is illuminated, is negative. Stimulated by the investigations of Hertz, Hallwachs made a more thorough investigation of the phenomenon, and his work, coupled with that of Elster and Geitel, Lenard, and J. J. Thomson, led to Einstein's

fundamental concept. He pointed out that so far as the photoelectric effect is concerned, light can be regarded as being made up of individual particles, or light quanta, each containing an amount of energy equal to $h\nu$ where h is the constant of Planck and ν is the frequency of the light.

$$h(\nu - \nu_0) = \frac{1}{2}mv_{\max}^2 = eV$$

$$h\nu_0 = e\phi$$

In this formulation of the Einstein photoelectric equation, the maximum kinetic energy of the electron ejected from the material is a function of the frequency of the exciting light minus the frequency of a threshold characteristic for the material. m , e and v are the mass, charge and velocity of the emitted electron, and V is the potential difference required to reduce the velocity of the emitted electron to zero. The quantity $h\nu_0$ is usually called the work function, whereby the work function ϕ is defined as the work necessary to remove to infinity an electron from the lowest free electron state in the metal, W_0 . If the electron is taken from the Fermi level, its work function $\phi = W_0 - W_f$.

Einstein's law predicts that the emission current should be directly proportional to the intensity of the incident light, a statement which underwent numerous critical investigations and was found to be correct. It is also consistent with a three-step model of the photoelectric emission from a solid. In the first step, the photon is absorbed. Absorption may be at any distance from the surface, governed by the light absorption constant of the material. As the emitted electron may be ejected at a finite depth, it has to diffuse through a crystal lattice to the surface and may lose energy in the diffusion process. The third step is the escape of the electron from the solid over a potential barrier.

While the above model is consistent with the Einstein equation, this latter does not predict the entire behavior of the photoemission from a solid. In particular, it does not say anything about the temperature coefficient. A theory was, therefore, necessary, and a quantum mechanical theory was provided by Fowler. On the basis of the initial assumption, that all electrons with normal components of energy greater than ϕ would escape, he established a relationship between the photoelectric current I and the temperature T

$$I = \alpha AT^2 \varphi(x)$$

where α and A are constants and x is defined by

$$x = \frac{h(\nu - \nu_0)}{kT}$$

The function φ is an exponential series, whose numerical values have been tabulated. The Fowler equation allows a very convenient plotting of photoelectric emission results, and the typical Fowler curve is a presentation of logarithm I/T^2 plus a constant vs x . Incidentally, this presentation allows a very convenient and fast determination of the wavelength threshold value, which is equivalent to the work function for the material.

The detailed remarks of the last two paragraphs are essentially limited to the photoelectric effect in pure metals and the ejection of the electron into vacuum. Most work functions are relatively high, with a resulting threshold in the ultraviolet region. ϕ ranges in metals from 1.9 eV (7000Å) for Cs to about 5.4 eV (2300Å) for Pt. Quite a bit of effort was made for many years to create photoemitters with low work functions for practical applications in the visible part of the spectrum or even in the infrared. One group of substances, offering considerably lower values than those of the metals, are semi-conductors. While in a metal the photoelectrons originate in the conduction band relatively close to the Fermi level, in semi-conductors they originate in the valence band or in impurity states. If the latter are more favorably placed than the Fermi level, a corresponding reduction of the work function may result. For instance, in Ag-O-Cs ϕ equals 1.0 eV (12300Å). It was found that intermetallic compounds of the I-V or II-V type, such as Cs₃Sb or Na-K-Cs-Sb, offer great advantages.

Observations in the early days of photoelectric research have shown that surface conditions may modify considerably the photoemissive properties of surfaces. The most commonly occurring surface impurity is an oxide layer, and for practical applications fortunately it was found that oxidation often enhances the photoelectric current. Although the mechanism of the photoemission of such surfaces is sometimes very complicated, part of the reason can be found in the enhanced light absorption (reduced reflection) of the photoelectric surface.

Practical application of surface photoemission in the early days was largely limited by the very weak currents which can be obtained from substances with a low efficiency of photoelectric emission. This led to schemes of charge multiplication which produce higher currents. A first such scheme used collision-ionization in a gas. Within a reasonable limit, the so-called gas-filled photocells show a very linear behavior with intensity. In the last 20 years, gas-filled photocells have been almost entirely replaced by photomultipliers, i.e., combinations of photocathodes with secondary electron multipliers. In these latter devices, the multiplication factor may vary from 10^4 to 10^7 .

With increasing attention to the far-ultraviolet range of the spectrum, a few metals with higher work functions find favor as applied devices. In some modern applications, advantage is taken of the insensitivity of a high work function photocathode to visible light. Its response is limited to far-ultraviolet light alone, in the so-called solar-blind type cell.

The second important group of phenomena, which has found many important practical applications, is that in which the ejected electron may travel through the emitting material, to enter a solid electrode in contact with the photoemitter instead of going to the anode through a vacuum. There are two subgroups: one is photoconductivity which is treated elsewhere in this volume, and the second is a group that bears the generic

name of photovoltaic effects. This name is generally applied to the phenomena leading to the direct conversion of a part of the energy absorbed from the impinging photons into usable electrical energy. Most common examples are selenium cells, the barrier-layer-type copper oxide cells, and the so-called solar batteries or solar photovoltaic converter, consisting usually of a semiconductor crystal like silicon containing a composition gradient termed the *p-n* junction. These last devices have now reached a relatively high grade of efficiency where up to 15 per cent of the energy of the incident light may be converted into electric current. Theoretically, 50 per cent efficiency might be obtained in a three-semiconductor sandwich assembly.

Photoelectric effects in liquids and in gases are essentially of the photoconductive type.

L. MARTON

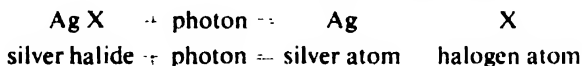
References

- Hughes, A. L., and duBridge, L. A., "Photoelectric Phenomena," New York, McGraw-Hill Book Co., 1932.
- Simon, H., and Suhrmann, R., "Der Lichtelektrische Effekt," Second edition, Berlin, Springer, 1958.
- Weissler, G., "Photo Ionization in Gases and Photoelectric Emission from Solids," in Flüge, S., Ed., "Encyclopedia of Physics," Vol XXI, pp. 304-382, Berlin, Springer, 1956.
- Gorlich, P., "Recent Advances in Photoemission," in Marton, L., Ed., *Advan. Electron. Electron Phys.*, **11**, 1-30 (1959).

Cross-references: PHOTOCONDUCTIVITY, PHOTOMULTIPLIER TUBE, PHOTON, PHOTOVOLTAIC EFFECT.

PHOTOGRAPHY

The history of photography spans over more than a century and embraces many techniques for producing photographic images. Within the past year or two alone, a number of important innovations have been made in photographic recording.



Although there is great diversity in image-forming optical systems, photosensitive materials, and development procedures, some form of each is essential to every photographic process. Light from the subject passes through the optical system or camera, and impinges on the photosensitive material to form a latent or invisible image. The latent image is transformed by chemical or physical means into a visible negative or positive image.

Most photographic processes use mixtures of silver halides (silver chloride, bromide, and iodide) as the photosensitive materials. These processes include conventional black and white photography and color photography, "Polaroid" black and white photography, and such recent innova-

tions as "Polacolor," three-dimensional color photography, and lensless photography. A number of other photographic processes are based on photochemical reactions of organic dyes and of inorganic metal salts other than silver halides. Still other processes employ light-induced changes in the physical properties of materials such as photoconductors or thermoplastics.

In conventional silver halide photography, the camera consists of a lightproof box with a lens for admitting and focusing the light, a diaphragm for controlling the size of the effective lens aperture, a shutter for regulating the length of exposure and the photosensitive film.

Two recent developments in photography are based on unconventional image-forming systems. In lensless photography¹, monochromatic light from a laser is split into two beams, one of which enters the lensless camera directly, while the other enters after passing through a transparent image. The two beams combine to form a Fresnel diffraction pattern of the image which is recorded on film. This recorded pattern is called a hologram. To reconstruct the image, a beam of laser light is transmitted through the hologram to generate a second diffraction pattern, a component of which reproduces the original image onto the screen.

The other new process is for mass producing a "three-dimensional" photograph known as a parallax panoramagram on a two-dimensional card.² The picture is coated with a plastic layer composed of vertical rows of cylindrical lenses which give the illusion of depth.

Ordinary photographic film is a suspension of microscopic silver halide crystals³ in gelatin coated on a supporting layer. Gelatin is a stable porous medium which allows the processing solutions to reach the dispersed crystals. Gelatin contains impurities which produce sensitivity centers. These are tiny silver sulfide centers of at least ten molecules each, which form on the surface of the crystals and catalyze the formation of the latent image by light.³

When the film is exposed, the following reaction takes place:

The latent image consists of small specks of metallic silver on the silver halide crystals. The smallest latent image speck that can make an exposed crystal develop during processing has from four to ten atoms. Not all absorbed photons contribute to this latent image. Assuming that a typical speck of ten atoms formed by absorbing one hundred photons can cause development of a crystal with 10^{10} ion pairs, the amplification factor is about 10^8 , which is higher than that of any other photographic material.

The developer reduces silver ions in the exposed crystals to silver atoms which deposit on the specks. The unexposed crystals are removed from the emulsion by a "fixing" solution to form the negative. A positive print is obtained by

exposing an emulsion-coated paper to white light through the negative and developing it.

In the "Polaroid" process developed by E. H. Land,³ a positive print is obtained directly from the camera one minute after exposure. The negative emulsion layer is pressed against a non-sensitive layer of paper between metal rollers in the camera. Sacs containing a viscous developer are ruptured as the two strips are pressed together, squeezing the developer uniformly between the strips. The developer develops the negative image, then dissolves the unexposed crystals and transfers them to the paper layer which contains dry developer to develop the positive print.

Silver halide emulsions are sensitive to ultraviolet and visible light with the maximum response in the blue at about 4200Å. Their range can be extended to any wavelength up to 12 000Å in the infrared by sensitizing dyes which absorb light in the desired spectral regions and transfer the energy to the crystals. The latent image is then formed in the usual way. Two common sensitized emulsions are panchromatic film (6200Å), and orthochromatic film (5600Å).

In modern color photography, integral tripack films are usually used. An integral tripack is a film of three sensitive layers separated by gelatin. Each layer contains crystals sensitized to one of the primary colors for light mixing, i.e., red, green, or blue. After exposure the three silver negatives in the nonseparable multilayered film are developed.

The most common method of forming the dye image is the dye coupling process. During development of the silver negative, the developer is oxidized in proportion to the silver reduced. The oxidized developer reacts or "couples" with a chemical to form a dye which produces a negative color image. A cyan dye in the red sensitized negative absorbs red light, while the blue sensitized emulsion acquires a yellow dye and the green sensitized emulsion a magenta dye. The superposition of the three color negatives then transmits the colors complementary to those of the original image.

To make color prints the colors are reversed by exposing a three layer emulsion on paper to white light through the color negative. To make positive transparencies the color reversal is performed directly on the three emulsion layers by changing the development procedure.

The remarkable "Polacolor" process recently developed by E. H. Land produces a color print fifty seconds after exposure¹ instead of the hour and a half required for the conventional process. The camera and the mechanical operations are similar to those for "Polaroid" black and white photography.

One innovation in this process is a new type of molecule with a preformed dye on one end and a developer on the other. The dye-developer linked molecules are placed inside the negative which is an integral tripack. After exposure the viscous reagent liberated from the pods sets the linked molecules in motion. Those that hit an exposed crystal in the appropriately sensitized emulsion

become fixed to the crystal site during development and form a negative color image. The other linked molecules move until they enter the positive which is pressed against the negative in the camera. The positive is a silver halide emulsion exposed together with the negative. The linked molecules containing the three types of dyes develop the exposed crystals to form the positive color print.

Photographic processes using photochemical systems other than silver halides have lower light sensitivity, but may be used to achieve special effects.⁵

A class of processes known as electrophotography uses materials which show an increase in electrical conductivity during light exposure. Electrophotography is widely used for office copying machines and for continuous tone photography.

In one type of electrophotography, charges are sprayed on a selenium or zinc oxide layer by a corona wire charged to five or ten thousand volts.³ During exposure, the charges leak off imagewise through the photoconductive layer. The remaining charge pattern is the latent image which is developed by applying an oppositely charged black resin powder which adheres to the charged areas. The image is fixed by slight heating to fuse the resin. "Electrofax" and "Xerography" are commercial adaptations of this process.

A type of electrophotography based on the "persistent internal polarization" effect in photoconductors promises to be even more useful.⁶ A potential of a hundred volts is applied across a zinc sulfide layer during exposure. Electrons in exposed regions are pulled toward the positive electrode and trapped en route by impurity atoms. The latent image is the imagewise internal polarization due to separation of trapped negative and positive charges. It is developed with the charged resin. This process has not yet been commercially exploited, although a few office copying machines have been built. However, it has several advantages over the other type of electrophotography, since the charging potential is much lower and the internal charges persist for a very long time. Negative or positive images may be made with equal ease. This process has been used to record wavelengths up to 20,000 or 30,000Å in the infrared where silver halides are not sensitive.

Thus, the past few years have brought about important innovations in silver halide and other photographic processes, and by extrapolation the future promises even greater wonders.

MIRIAM SIDRAN

References

1. Leith, E., and Upatnieks, J., "Wavefront Reconstruction with Continuous-Tone Objects," *J. Opt. Soc. Am.*, **53** 1377 (1963).
2. Lipton, L., "Color 3-D Printing Process Permits Mass Press Run, Glassless Viewing," *Popular Phot.*, 92 (May 1964).

3. Neblette, C. B., "Photography, Its Materials and Processes," Princeton, N.J., D. Van Nostrand Company, Inc., Sixth edition, 1962.
4. "Background Information about Polaroid Land Color Film," Polaroid Corporation, 1962.
5. Hefner, M., "Not by Silver Alone," *Perspective*, 1, 28 (1959).
6. Kallmann, H., Rennert, J., and Sidran, M., "A Photographic Process using Persistent Internal Polarization in Phosphors," *Phot. Sci. Eng.*, 4, 345 (1960).

Cross-references: COLOR, LENS, OPTICAL INSTRUMENTS, PHOTOCONDUCTIVITY

PHOTOMETRY

Photometry is concerned with the measurement of light, or much more frequently, of the time rate of flow of light. Light is the aspect of radiant energy to which the human eye responds, or more precisely it is Q_v in the equation:

$$Q_v = K_m \int V(\lambda) Q_{e,\lambda} d\lambda$$

in which K_m is the maximum luminous efficacy of radiant energy (about 680 lumens/watt), $V(\lambda)$ is the CIE (International Commission on Illumination) spectral luminous efficiency function¹ which is curve 4 of the figure shown in the article on COLOR in this Encyclopedia, and $Q_{e,\lambda}$ is the spectral distribution of radiant energy.

The quantities with which photometry is concerned are luminous intensity and flux, illuminance and LUMINANCE, and luminous reflectance and transmittance.

The devices which are used to measure these quantities are called photometers. They fall into two general categories, namely, visual and physical. The making of brightness matches is a requisite in the use of visual photometers, while in physical photometers, which are usually of the photoelectric or thermoelectric type, luminous energy incident on a receiver is transformed into electric energy and the latter is measured by means of sensitive electric meters or devices. In visual photometry, observers with so-called normal color vision are used, it being tacitly assumed that the average spectral response of the observers used to make the observations approximates the standard CIE luminous efficiency curve. In physical photometry, the receiver should be equipped with a filter, the spectral transmittance of which when multiplied by the relative spectral response of the receiver, wavelength by wavelength, closely approximates the CIE luminous efficiency curve. In all photometric measurements which involve differences between the spectral distribution of the standard source used to calibrate the photometer and that of the test source, the accuracy of the results obtained is dependent on the degree of approximation of the spectral response curve of the photometer to the CIE spectral luminous efficiency curve.

The unit of luminous intensity (often called candlepower) is the candela. Its magnitude is

such that the luminous intensity of one-sixtieth of one square centimeter of projected area of a blackbody at the freezing point of platinum² is one candela. Such a blackbody has been the international standard of light since January 1, 1948. National standardizing laboratories, like the National Bureau of Standards in the United States, calibrate incandescent-lamp reference standards of luminous intensity in terms of the international standard. The calibration for luminous intensity of other incandescent lamps relative to these reference standards or to other standards of luminous intensity is usually done on a bar photometer, which consists of a horizontal bar equipped with movable carriages, on one of which the standards and lamps to be calibrated are mounted in turn, while the photometric measuring device is mounted on the other carriage. The calibration involves the use of the inverse-square law which states that the illuminance at a point on a surface varies directly with the luminous intensity I of the source and inversely as the square of the distance d between the source and the point on the surface. If the surface at the point is perpendicular to the direction of the incident light, the law may be expressed as follows: $E \propto I/d^2$. In the measurement of the intensity I_2 of a source relative to the intensity I_1 of a standard, the illumination at the photometer from the two sources is made equal by varying the distance d , so that $I_2 = I_1 d_2^2/d_1^2$ in which d_2 and d_1 are the distances for the unknown and standard, respectively. The important precautions which must be taken in calibrating and using standards of luminous intensity are (1) that the distance between the standard and photometric receiver is sufficiently large relative to the size of the source and of the receiver so that the source and receiver act approximately as points,³ (2) that the standard is accurately oriented with respect to the receiver,⁴ and (3) that by the use of baffles no flux is incident on the receiver other than that which comes directly from the standard.

Illuminance is incident luminous flux per unit area. In the measurement of illuminance, a photometer, in this application often called an illuminometer, is used. The illuminometer is calibrated at n points by the use of a standard of luminous intensity I placed at n distances d from the test plate of a visual photometer or the surface of the receiver, thus yielding n illuminances E computed from the inverse-square law $E = I/d^2$. It is important that the geometric characteristics of the receiver of the illuminometer be such as to enable it to evaluate, without bias, luminous flux which is incident on the receiver from all directions; special precautions are usually necessary if flux incident at large angles to the perpendicular to the receiver is to be properly evaluated.

The unit of luminous flux is the lumen which is defined as the luminous flux in a unit solid angle (steradian) from a point source having a uniform intensity of one candela. Standards of luminous flux are calibrated in terms of standards of luminous intensity by the use of a two-step procedure. In one step, the luminous intensity in a specified

direction of the lamp to be calibrated is determined, and in the other step, by the use of a distribution photometer, the relative luminous intensity of this lamp in a multiplicity of directions is determined. To enable one to see how each of these intensities is effective in contributing luminous flux, one notes that each direction intersects a hypothetical sphere circumscribed about the lamp and that around the point of intersection on the sphere surface an area can logically be assigned which defines the solid angle in which the luminous intensity in that direction is effective in supplying luminous flux. These solid angles in steradians, by which the luminous intensity values must be multiplied to obtain the lumens incident on the respective areas, are called zonal constants. Such constants for various patterns of distribution measurements have been published and are thus readily available.^{5,6} Reference standards of luminous flux are calibrated by this procedure, but other standards are calibrated in terms of the reference standards or of other standards of luminous flux by the use of an integrating sphere, sometimes called an Ulbricht sphere. It is a spherical hollow enclosure with a uniform, diffusely reflecting inner surface. In such an enclosure, the illuminance due to reflected flux is the same at every point on the surface and is directly proportional to the flux emitted by a source in the sphere, independent of its angular distribution. The ratio of the fluxes emitted by two sources is thus the ratio of the illuminances at any point on the sphere surface which is shielded from receiving flux directly from the sources. The principal precautions in the use of spheres are (1) that the inner surface be a good diffuser, uniform in reflectance from point to point, and (2) that the spectral selectivity of the sphere throughout the visible region of the spectrum (0.4 to 0.7 μ) be compensated. This compensation is usually accomplished by the use of filters, but another procedure is to disperse the flux from the sphere wall by means of a prism and insert in the spectrum thus formed a template whose shape compensates for both the spectral selectivity of the sphere and the spectral response of the receiver.⁷ It must be remembered that the effect of the spectral selectivity is greatly amplified because of the interreflections within the sphere so that the illuminance on the sphere wall by reflection only is $\rho/(1 - \rho)$ times the average illuminance by directly incident flux, where ρ is the reflectance of the sphere wall; thus if the reflectances at two wavelengths are, for example, 0.80 and 0.72, the spectral irradiance by reflection only at the wavelength of lower reflectance will be only 64 per cent of that at the wavelength of higher reflectance.

There are two related systems of units of luminance which have been and are still being used. In one of these systems the luminance is expressed in terms of luminous intensity per unit projected area or in terms of luminous flux per unit solid angle and unit projected area. The other system is that in which luminance is expressed in terms of the flux per unit area which would leave a

perfect diffuser having the same luminance; the lambert and footlambert are units of this system and their magnitudes are such that one lumen per square centimeter and per square foot, respectively, would leave a perfect diffuser of unit luminance. Standards of luminance are usually combinations of luminous intensity standards and surfaces of known directional transmittance factor or directional reflectance factor (see below). The product of the illuminance of a surface by the appropriate one of these factors gives its luminance; for example, for each lumen per square centimeter incident perpendicularly on a freshly prepared surface of MgO, the luminance at an angle of 45° from the perpendicular will be approximately one lambert because the directional reflectance factor of MgO for this geometry is approximately 1.00. In lieu of the MgO surface, use is more generally made of a highly diffusing glass or plastic plate of known directional transmittance factor for perpendicular incidence on one side of the plate and perpendicular viewing on the other side. The measurement of luminance relative to a luminance standard by visual photometry introduces no serious problems other than those inherent in photometry generally, because the receptor, in this case the eye, intercepts flux in a very small solid angle which can, for all practical purposes, be considered to be infinitesimally small. In physical photometry, however, because of the need to collect flux in an amount which will result in adequate sensitivity, the solid angle of reception is relatively large and cannot in general be considered to be infinitesimally small, so that there is introduced into the measurements an inaccuracy whose magnitude is related to the magnitude of the variation of the flux per infinitesimal solid angle within the solid angle of reception.

In addition to the measurement of the dimensional quantities discussed above, photometry is concerned with the measurement of the dimensionless quantities transmittance, directional transmittance factor, reflectance, and directional reflectance factor. Transmittance and reflectance are the ratio of transmitted and reflected flux, respectively, to incident flux. For a nondiffusing specimen, the measurement of transmittance or reflectance usually poses no great problem. For a diffusing specimen, the measurement of totally diffuse transmittance or reflectance (reception over a solid angle of 2π steradians) for specified modes of illumination of the specimen is usually made with instruments that incorporate integrating spheres;⁸ for solid angles of reception smaller than 2π steradians, the measurement of the ratio of transmitted or reflected flux to incident flux yields what is designated as fractional transmittance or fractional reflectance, respectively, quantities which generally are not of as much interest as are the directional transmittance factor and the directional reflectance factor (usually in the United States widely but inappropriately called directional transmittance and directional reflectance). These factors are defined as the ratio of the flux transmitted (or

TABLE 1. PHOTOMETRIC AND RELATED RADIOMETRIC QUANTITIES, DEFINING EQUATIONS, UNITS AND SYMBOLS

Quantity*	Symbol*	Defining Equation	Unit	Symbol
Radiant energy	Q		Erg Joule Calorie Kilowatt-hour	J cal kWh
Radiant density	w	$w = dQ/dV$	Joule per cubic meter Erg per cubic centimeter	J/m^3 erg/cm^3
Radiant flux	Φ	$\Phi = dQ/dt$	Erg per second Watt	erg/s W
Radiant flux density at a surface				
Radiant emittance** (Radiant exitance)	M	$M = d\Phi/dA$	Watt per square centimeter	W/cm^2
Irradiance	E	$E = d\Phi/dA$	Watt per square meter, etc.	W/m^2
Radiant intensity	I	$I = d\Phi/d\omega$ (ω = solid angle through which flux from point source is radiated)	Watt per steradian	W/sr
Radiance	L	$L = d^2\Phi/d\omega(dA \cos \theta)$ $= dI/(dA \cos \theta)$ (θ = angle between line of sight and normal to surface considered)	Watt per steradian and square centimeter Watt per steradian and square meter	$W \cdot sr^{-1} cm^{-2}$ $W \cdot sr^{-1} m^{-2}$
Emissivity	ϵ	$\epsilon = M/M_{blackbody}$	None (dimensionless)	—
Absorptance	α	$\alpha = \Phi_a/\Phi_i$ ***	None (dimensionless)	—
Reflectance	ρ	$\rho = \Phi_r/\Phi_i$ ***	None (dimensionless)	—
Transmittance	τ	$\tau = \Phi_t/\Phi_i$ ***	None (dimensionless)	—
NOTE: The symbols for photometric quantities (see below) are the same as those for the corresponding radiometric quantities (see above). When it is necessary to differentiate them the subscripts v and e respectively should be used, e.g., Q_v and Q_e .				
Luminous energy (quantity of light)	Q	$Q_v = \int_{380}^{760} K(\lambda)Q_e d\lambda$	Lumen-hour	$lm \cdot h$
Luminous density	w	$w = dQ/dV$	Lumen-second (talbot) Lumen-second per cubic meter	$lm \cdot s$ $lm \cdot s \cdot m^{-3}$
Luminous flux	Φ	$\Phi = dQ/dt$	Lumen	lm
Luminous flux density at a surface				
Luminous emittance† (Luminous exitance)	M	$M = d\Phi/dA$	Lumen per square foot	lm/ft^2
Illumination (Illuminance)	E	$E = d\Phi/dA$	Footcandle (lumen per square foot) Lux (lm/m^2) Phot (lm/cm^2)	fc lx ph
Luminous intensity (candlepower)	I	$I = d\Phi/d\omega$ (ω = solid angle through which flux from point source is radiated)	Candela (lumen per steradian)	cd
Luminance (photometric brightness)	L	$L = d^2\Phi/d\omega(dA \cos \theta)$ $= dI/(dA \cos \theta)$ (θ = angle between line of sight and normal to surface considered)	Candela per unit area Stilb (cd/cm^2) Nit (cd/m^2) Footlambert ($cd/\pi ft^2$) Lambert ($cd/\pi cm^2$) Apostilb ($cd/\pi m^2$)	cd/m^2 , etc. sb nt fL L asb
Luminous efficiency	V	$V = K/K_{max}$	None (dimensionless)	—
Luminous efficacy	K	$K = \Phi_v/\Phi_e$	Lumen per watt	lm/W

* Quantities may be restricted to a narrow wavelength band by adding the word spectral and indicating the wavelength. The corresponding symbols are changed by adding a subscript λ , e.g., Q_λ , for a spectral concentration or λ in parentheses, e.g., $K(\lambda)$, for a function of wavelength.
 ** Should be deprecated in favor of emitted radiant exitance.
 † Should be deprecated in favor of emitted luminous exitance.
 *** Φ_i = incident flux; Φ_a = absorbed flux; Φ_r = reflected flux; Φ_t = transmitted flux.

reflected) in the solid angle of interest to that which would be transmitted (or reflected) by the ideal perfect diffuse transmitter (or reflector) identically illuminated; the ideal perfect diffuse transmitter (or reflector) is one which transmits (or reflects) all of the luminous flux incident on it in accord with the Lambert cosine law, i.e., so that the flux per unit solid angle in any direction from it varies as the cosine of the angle between that direction and the perpendicular to the transmitter (or reflector). For a solid angle of reception of 2π steradians, the term directional transmittance (or reflectance) factor is synonymous with transmittance (or reflectance). For infinitesimal solid angles of reception and nonfluorescing specimens, the term directional transmittance (or reflectance) factor is synonymous with luminance factor which for any specimen is defined as the ratio of the luminance of the specimen to the luminance of a perfect diffuser identically illuminated.

The most commonly used photometric and related radiometric quantities, their defining equations and units, and symbols for them which are consistent with those agreed upon to date by the International Commission on Illumination, the International Electrotechnical Commission, the International Standards Organization, and the Commission for Symbols, Units, and Nomenclature of the International Union of Pure and Applied Physics are listed in Table 1.

L. E. BARBROW

References

1. *Proceedings of the International Commission on Illumination* (1924).
2. Wensel, H. T., Roeser, W. F., Barbrow, L. E., and Caldwell, F. R., "The Waidner-Burgess Standard of Light," *J. Res. Natl Bur. Std.*, **6**, 1103 (June 1931).
3. Walsh, J. W. T., "Photometry," 3rd Ed., New York, Dover Publications, 1958.
4. Barbrow, L. E., Wilson, S. W., "Vertical Distribution of Light from Gas-Filled Candlepower Standards," *Illum. Eng.*, **53**, 645 (December 1958).
5. "IES Lighting Handbook," Illuminating Engineering Society, New York, N.Y.

6. Cotton, H., "Principles of Illumination," New York, John Wiley & Sons, 1961.
7. Winch, G. T., "Photometry and Colorimetry of Fluorescent and Other Electric Discharge Lamps," *Trans. Illum. Eng. Soc.* (London), **11**, 107 (June 1946).
8. Taylor, A. H., "Errors in Reflectometry," *J. Opt. Soc. Am.*, **25**, 51 (February 1935).

Cross-references: COLOR; LUMINANCE; MEASUREMENTS PRINCIPLES OF; OPTICS; GEOMETRICAL; REFLECTION.

PHOTOMULTIPLIER

Photomultipliers make use of the phenomena of photoemission and secondary-electron emission in order to detect very low light levels. The electrons released from the photocathode by incident light are accelerated and focused onto a secondary-emission surface (called a dynode). Several electrons are emitted from the dynode for each incident primary electron. These secondary electrons are then directed onto a second dynode where more electrons are released. The whole process is repeated a number of times depending on the number of dynodes. In this manner, it is possible to amplify the initial photocurrent by a factor of 10^8 or more in practical photomultipliers. It is, therefore, evident that the photomultiplier represents an extremely sensitive detector of light.

The major characteristics of the photomultiplier with which the user is generally most concerned are as follows:

- (a) sensitivity, spectral response, and thermal emission of photocathodes;
- (b) amplification factor;
- (c) noise characteristics and the signal-to-noise ratio.

Sensitivity, Spectral Response, and Thermal Emission of Photocathodes. Many different types of photocathodes are being used in photomultipliers. With a selection of various cathodes, it is possible to cover the range of response from the soft x-ray region (approximately 5 to 500Å) to the near infrared (approximately 12,000Å). Photocathodes in common use are listed in Table 1. Typical spectral response curves are shown in

TABLE 1. CHARACTERISTICS OF COMMON PHOTOCATHODES

Photocathode	Retma Code Number	λ_m , at max. (Å)	λ_0 , 1% of max. (Å)	Quantum Efficiency, at λ_m	Average Thermionic Emission at 25°C (amp/cm ²)	Average $\mu A/lumen$ (2870°K tungsten source)
Cs-O-Ag	S-1	8000	12,000	0.005	10^{-11}	10-20
Cs-Sb(0) opaque	S-4	4000	7,000	0.25	10^{-14}	60-100
Cs-Sb(0)	S-11	4500	7,000	0.20	10^{-14}	40-80
Cs-Ag-Bi	S-10	4500	7,500	0.10	10^{-13}	40-60
Na-K-Sb		4000	6,200	0.20	10^{-16}	30-60
Na-K-Cs-Sb	S-20	4250	8,250	0.25	10^{-14}	150-180
CuI		1100	1,900	0.30		
CsI		1250	1,950	0.30		

Fig. 1. The thermal emission at 25°C of CuI and CsI is lower than that of the other cathodes listed, however no quantitative measurements have as yet been published for them.

Amplification Factor. The amplification factor in a photomultiplier depends on the secondary emission characteristics of the dynode and to some extent on the design of the multiplier structure.

Most secondary-emission surfaces used in commercial photomultipliers fall into two classes:

(1) Alkali metal compounds, e.g., cesium antimony.

(2) Metal oxide layers, e.g., magnesium oxide on silver-magnesium alloy.

The alkali metal compounds have higher gain at low primary electron energy (of the order of 75 V). The metal oxide layers show less fatigue at high current density of emission (i.e., at several microamperes per square centimeter or higher).

Table 2 lists some characteristics of the common dynode surfaces.

TABLE 2. CHARACTERISTICS OF COMMON DYNODE SURFACES

Surface	Maximum Secondary Emission Ratio	Primary Voltage for Maximum Ratio
Cs-Sb	8.0	500
Cs-Ag-O	5.8-9.5	500-1000
MgO (on AgMg)	9.8	500
BeO (on CuBe)	3.5-5.5	500-700
BeO (on NiBe)	12.3	700
Al ₂ O ₃	1.5-4.8	350-1300

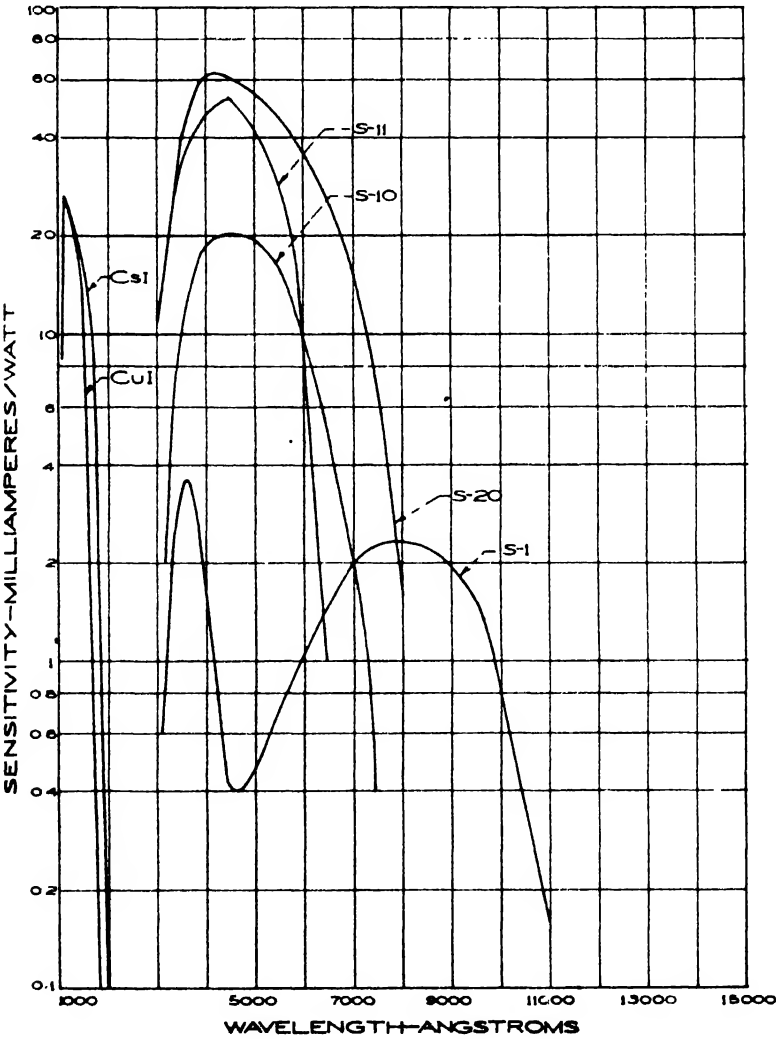


FIG. 1. Spectral response characteristics.

The multiplier structures may be divided into two main types, dynamic and static.

The dynamic multiplier in its simplest form consists of two parallel dynode surfaces with an alternating electric field applied between them. Electrons leaving one surface at the proper phase of the applied field are accelerated to the other surface where they knock out secondary electrons. These electrons in turn are accelerated back to the first plate when the field reverses, creating still more secondary electrons. Eventually the secondary electrons are collected by an anode placed in the tube; if they are not, a self-maintained discharge occurs. In practice, dynamic multipliers have been replaced by static ones mainly because the latter have better stability and are easier to operate.

The static multipliers may be either magnetically or electrostatically focused.

Figure 2(A) illustrates one type of multiplier structure using *magnetic* focusing. Primary electrons impinging on one side of a dynode cause the emission of secondary electrons from the opposite side. These electrons are then focused onto the next dynode by means of the axial magnetic field.

Figures 2(B) through (F) illustrate the more

common types of *electrostatic* multiplier structures. The structures shown in Fig. 2(B), (C), and (D) actually use focusing from one stage to the next. The structures in Fig. 2(E) and (F) are unfocused.

Recent technology has made possible the deposition of thin semiconductor secondary emission surfaces onto insulating substrates. Such dynode strips have been used in designing rugged miniature multiplier structures. Two such structures are shown in Fig. 2(G) and (H). In both these structures the secondary electrons are accelerated along the length of the dynode strip by means of the potential applied across the strip. The structure in Fig. 2(H) can be made particularly small, the dimensions of a typical unit being 0.5 mm outside diameter and 10 mm long.

The unfocused electrostatic structures have less sensitivity to stray electric and magnetic fields. The focused structures, especially (C) and (D) in Fig. 2, can be made to have very short transit-time spreads. Often special accelerating electrodes are placed between the dynodes to improve their transit-time spread and space-charge characteristics. At 200 to 300 V/stage, the transit time spread may be less than 10^{-9} second, and peak

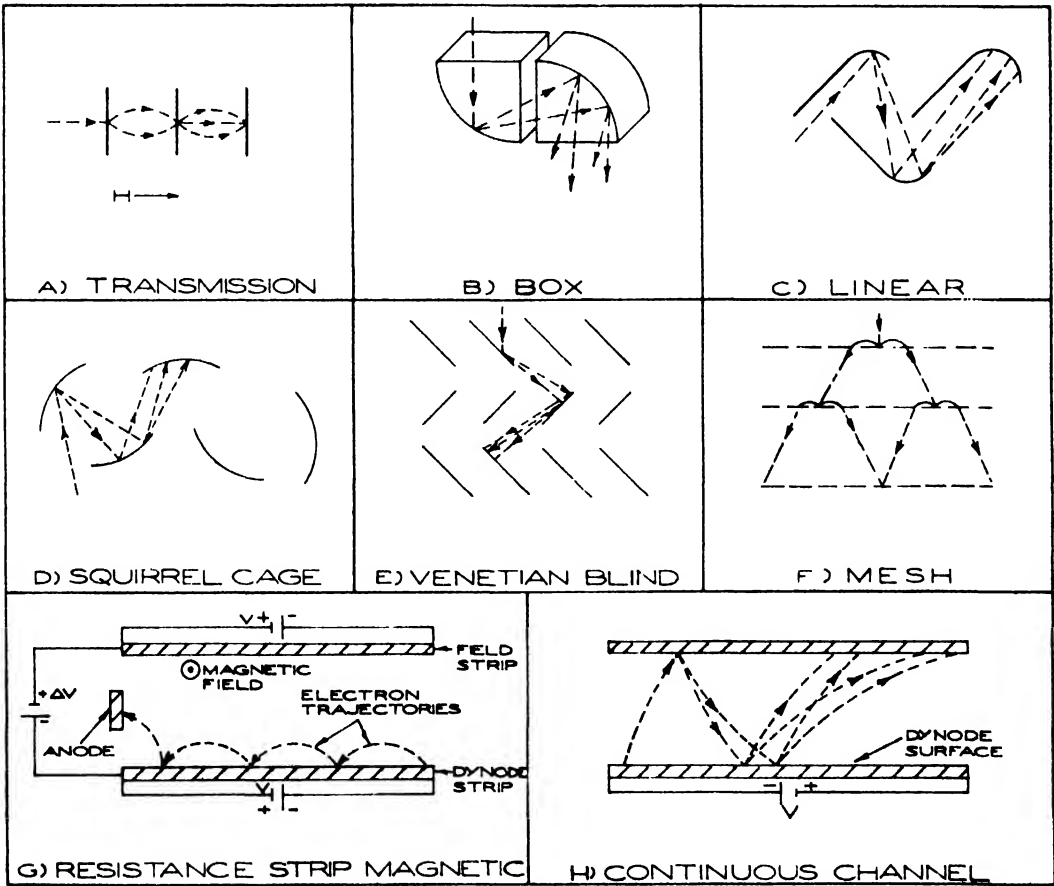


FIG. 2. Various multiplier structures.

pulse outputs of the order of 200 mA may be drawn before nonlinearity owing to space-charge saturation sets in.

In the normal operating range, the over-all amplification G of the multiplier is proportional to V^β , where V is the over-all voltage and β is a constant of the order of seven for a ten-stage structure.

Noise Characteristics and the Signal-to-noise Ratio. It is necessary to distinguish between two types of noise in photomultipliers, dark and shot noise.

The *dark noise* in photomultipliers is caused mainly by the following:

- (1) leakage current across insulating supports,
- (2) field emission from electrodes,
- (3) thermal emission from the photocathode and dynodes,
- (4) positive ion feedback to the photocathode, and
- (5) fluorescence from dynodes and insulator supports.

By careful design and construction of the photomultiplier it is possible to limit the dark noise principally to item (3).

Associated with the photocurrent from the photocathode is *shot noise*. There is also shot noise from secondary emission in the multiplier structure. The mean-square shot noise current $\overline{i_n^2}$ at the anode is given to a close approximation by

$$\overline{i_n^2} = 2ebG^2\overline{i_p}\Delta f$$

where e is the electronic charge, G is the amplification, $\overline{i_p}$ is the average photocurrent, Δf is the bandwidth of the system, and b is a factor equal to approximately 1.5 which accounts for the shot noise created in the multiplier. The signal-to-noise ratio S/N is then given by

$$S/N = \frac{G\overline{i_p}}{\sqrt{\overline{i_n^2}}} = \sqrt{\frac{\overline{i_p}}{2eb\Delta f}}$$

Uses of Photomultipliers. One of the major uses of photomultipliers is in the scintillation counter where in combination with a fluorescent material it is used to detect nuclear radiation. The advent of the space program has led to applications in star and planet tracking for guidance systems as well as in star photometry and quantitative measurements of soft x-rays in outer space. Other applications include use in facsimile transmission, spectral analysis, automatic process control, and many other areas where extremely low light levels must be detected.

BERNARD R. LINDEN

References

- Engstrom, R. W., "Multiplier Phototube Characteristics. Application to Low Light Levels," *J. Opt. Soc. Am.*, **37**, 420-430 (1947).
 Dunkelman, L., et. al., "Spectrally Selective Photodetectors for the Middle and Vacuum Ultraviolet," *Appl. Opt.*, **1**, 695-700 (1962).

Linden, B. R., "Five New Photomultipliers for Scintillation Counting," *Nucleonics*, **11**, 30-33 (1953).

Rodman, H. P., and Smith, H. J., "Tests of Photomultipliers for Astronomical Pulse Counting Applications," *Appl. Opt.*, **2**, 181-186 (1963).

See also the Proceedings of the Scintillation Counting Symposia in *IRE Trans. Nucl. Sci.*, **3**, No. 4, 1956; **5**, No. 3, 1958; **7**, No. 2-3, 1960; **9**, No. 3, 1962.

Cross-references: PHOTOCONDUCTIVITY; PHOTOELECTRICITY.

PHOTON

The photon is the quantum of the electromagnetic field. It is a particle with zero rest mass and spin one. For a photon moving in a specific direction, the energy E and momentum p of the particle are related to the frequency f and wavelength λ of the field by the Planck equation $E = hf$ and the de Broglie equation $p = h/\lambda$. As for all massless particles, the energy and momentum are related by $E = cp$ and the photon can only exist moving at light velocity c . Another property of all massless particles is this: given the momentum, the particle can exist in just two states of spin orientation. The spin can be parallel or antiparallel to the momentum but no other directions are possible. The photon state with the spin and momentum parallel/antiparallel is said to be right-/left-handed and is a right-/left-hand circularly polarized wave. In analogy with the neutrino, one can say that the state has positive/negative helicity and can call the right-handed particle the antiphoton, the left-handed particle the photon. There is an operation, CP conjugation, that converts a photon state into an antiphoton state and vice versa. It is possible to superpose photon and antiphoton states in such a way that the superposition is unchanged by CP conjugation and so gives a type of photon that is its own antiparticle. The photons produced by transitions between states of definite parities in atoms or nuclei are their own antiparticles in this sense. As for all particles with integer spin, the photon follows Bose-Einstein statistics. This means that a large number of photons may be accumulated into a single state. Macroscopically observable electromagnetic waves, such as those resonating in a microwave cavity for example, are understood to be large numbers of photons all in the same state. The photon, among all the particles, is unique in having its states be macroscopically observable this way.

The electric and magnetic fields \mathbf{E} and \mathbf{B} describe the state of the photon and make up the wave function of the particle. Maxwell's equations give the time development of the fields and take the place, for the photon, that Schrödinger's equation takes for a material particle. Many of the remarks above follow as direct consequences of Maxwell's equations. In Gaussian units, where both \mathbf{E} and \mathbf{B} are measured in gauss or dynes per

electrostatic unit of charge, the equations are

$$\epsilon_{jkl}\partial E_l/\partial x_k + c^{-1}\partial B_j/\partial t = 0 \quad (1)$$

$$\epsilon_{jkl}\partial B_l/\partial x_k - c^{-1}\partial E_j/\partial t = 0 \quad (2)$$

$$\partial E_j/\partial x_j = \partial B_j/\partial x_j = 0 \quad (3)$$

The particle aspect of the equations becomes evident when the equations are written in terms of the complex three-vector

$$\psi_j = E_j + iB_j \quad (4)$$

in which case they become

$$\epsilon_{jkl}\partial\psi_l/\partial x_k - ic^{-1}\partial\psi_j/\partial t = 0 \quad (5)$$

$$\partial\psi_j/\partial x_j = 0 \quad (6)$$

Equation (6) is to be considered as an initial condition rather than as an equation of motion since it follows from Eq. (5) that

$$c(\partial\psi_j/\partial x_j)/\partial t = -ic\epsilon_{jkl}\partial^2\psi_l/\partial x_j\partial x_k = 0$$

so if $\partial\psi_j/\partial x_j$ is zero at the start, it is zero forever. Equation (5) can be cast into Hamiltonian form. One writes the three components ψ_j as a column matrix ψ and introduces three, three-by-three matrices by

$$(s_k)_{jl} = i\epsilon_{jkl} \quad (7)$$

With this notation, Eq. (5) becomes

$$-ic(s_k)_{jl}\partial\psi_l/\partial x_k = i\partial\psi_j/\partial t$$

or

$$H\psi = i\hbar\partial\psi/\partial t \quad (8)$$

where

$$H = c\mathbf{s} \cdot \mathbf{p} \quad (9)$$

and \mathbf{p} is $-\hbar\nabla$. The Hamiltonian for the photon is thus $c\mathbf{s} \cdot \mathbf{p}$. In detail, the matrices that occur here are

$$s_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad s_2 = \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{pmatrix}, \quad (10)$$

$$s_3 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

They are Hermitian and, as is easily verified, they fulfil the commutation rules

$$[s_i, s_j] = i\epsilon_{ijk}s_k$$

and so are a set of angular momentum matrices. Evidently each has eigenvalues 0, ± 1 so they are a representation of spin one.

Next consider the plane wave solutions. Let them be propagating in the 3-direction; so substitute

$$\psi = u \exp [i\hbar^{-1}(p_3z - Wt)]$$

into Eq. (8). Here the same symbol p_3 is used for the eigenvalue as for the operator. This reduces

to the eigenvalue problem

$$c \begin{pmatrix} 0 & -ip_3 & 0 \\ ip_3 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} u = Wu$$

The eigenvalues are found to be $W = 0, \pm cp$, where p is $|p_3|$, and the corresponding eigenvectors are

$$u_0 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad u_{\pm} = \frac{1}{\sqrt{2}} \begin{pmatrix} \pm p_3/p \\ i \\ 0 \end{pmatrix}$$

The $W = 0$ possibility does not satisfy the initial condition, Eq. (6), and so must be discarded. The solutions u_{\pm} are valid for either sign of p_3 ; choose $p_3 = \pm p$ so both waves are propagating in the positive z direction. The two solutions of the problem are then

$$\psi_{\pm} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix} \exp [\pm i\hbar^{-1}p(z - ct)] \quad (11)$$

The subscript \pm denotes that these are eigenstates of the helicity operator $\mathbf{s} \cdot \mathbf{p}/p$ with eigenvalues ± 1 . The electric and magnetic fields are the real and imaginary parts:

$$E_{\pm, x} = 2^{-1} \cos [p\hbar^{-1}(z - ct)] \quad (12a)$$

$$E_{\pm, y} = \mp 2^{-1} \sin [p\hbar^{-1}(z - ct)] \quad (12b)$$

$$B_{\pm, x} = \pm 2^{-1} \sin [p\hbar^{-1}(z - ct)] \quad (12c)$$

$$B_{\pm, y} = 2^{-1} \cos [p\hbar^{-1}(z - ct)] \quad (12d)$$

$$E_z = B_z = 0 \quad (12e)$$

Here it is seen that the \pm helicity solution is right- left-hand circularly polarized with respect to the propagation direction.

The allowed states of the photon are eigenstates of the Hamiltonian H with eigenvalues $\pm cp$. Let $|H\rangle$ be the operator which, applied to the same states, gives eigenvalue cp . The operators for the physical energy, momentum, and angular momentum of the photon are $|H\rangle$, $(H/|H|)\mathbf{p}$, and $(H/|H|)(\mathbf{x} \times \mathbf{p} + \hbar\mathbf{s})$. One can understand these assignments for the energy and momentum by considering the plane wave states of Eq. (11). The states are eigenstates of the operators with energy eigenvalue cp and with momentum eigenvalue p in the positive z direction. As further justification for these operator assignments, the expectation values of the operators are directly related to the classical formulas for energy, momentum, and angular momentum in the electromagnetic field:

$$\int d^3x \phi^\dagger |H| \phi = (8\pi)^{-1} \int d^3x (E^2 + B^2) \quad (13a)$$

$$\int d^3x \phi^\dagger (H/|H|) \mathbf{p} \phi = (4\pi c)^{-1} \int d^3x (\mathbf{E} \times \mathbf{B}) \quad (13b)$$

$$\int d^3x \phi^\dagger (H/|H|) (\mathbf{x} \times \mathbf{p} + \hbar\mathbf{s}) \phi = (4\pi c)^{-1} \int d^3x \mathbf{x} \times (\mathbf{E} \times \mathbf{B}) \quad (13c)$$

where the function ϕ is defined by

$$\phi = |8\pi H|^{-1} \psi \quad (14)$$

These results apply for any solution ψ of Eqs. (6) and (8). The operation $|8\pi H|^{-1}$ in Eq. (14) is to be carried out by expanding ψ in plane wave components like ψ_+ and replacing $|8\pi H|^{-1}$ by $(8\pi c p)^{-1}$ in each component. Proofs of Eqs. (13) will not be given here; they can be made by expressing each side of the equations in terms of the plane wave expansion coefficients. Accepting these operator assignments, one sees that the helicity operator $\mathbf{s} \cdot \mathbf{p}$ is the component of the spin of the photon in the direction of its momentum.

The CP conjugation operation is related to the space reflection covariance of Maxwell's equations. Consider a primed and an unprimed coordinate system such that the coordinates of any point in space referred to the two axes are related by $\mathbf{x}' = -\mathbf{x}$. Suppose the electric field is axial and the magnetic field is polar so that the functions describing the fields are related by $\mathbf{E}'(\mathbf{x}', t) = \mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}'(\mathbf{x}', t) = -\mathbf{B}(\mathbf{x}, t)$. It is evident that Maxwell's equations have the same form in both coordinate systems and that the transformation rule for ψ is $\psi'(\mathbf{x}', t) = \psi^*(\mathbf{x}, t)$ where the asterisk denotes the complex conjugate. The fact that the equations have the same form in both systems implies further that if $\psi(\mathbf{x}, t)$ is any solution then $\psi'(\mathbf{x}, t)$, or equivalently $\psi^*(-\mathbf{x}, t)$, is also a solution. The operation that carries $\psi(\mathbf{x}, t)$ into $\psi'(\mathbf{x}, t)$ is called CP conjugation and one writes

$$\psi^{CP} = KP\psi \quad (15)$$

where K is the operation "take complex conjugate" and the operator P changes \mathbf{x} into $-\mathbf{x}$. If ψ is a solution of Maxwell's equations so also is ψ^{CP} . However KP anticommutes with $\mathbf{s} \cdot \mathbf{p}$ so if the solution ψ has \pm helicity, then ψ^{CP} has \mp helicity. The CP conjugation thus converts the particle into the antiparticle. The KP operator also anticommutes with the physical momentum operator so for a state $\psi_1(\mathbf{q})$ with definite helicity \pm and physical momentum \mathbf{q} one has

$$KP\psi_1(\mathbf{q}) = \psi_2(-\mathbf{q}) \quad (16)$$

Instead of the two states ψ_+ and ψ_- , one may consider the superpositions

$$\psi_1(\mathbf{q}) = 2^{-1}[\psi_-(\mathbf{q}) + \psi_+(\mathbf{q})] \quad (17a)$$

$$\psi_2(\mathbf{q}) = 2^{-1}[\psi_-(\mathbf{q}) - \psi_+(\mathbf{q})] \quad (17b)$$

The reason for introducing them is the property

$$KP\psi_1(\mathbf{q}) = \psi_1(-\mathbf{q}) \quad (18a)$$

$$KP\psi_2(\mathbf{q}) = -\psi_2(-\mathbf{q}) \quad (18b)$$

Thus the KP operation applied to ψ_1 or ψ_2 reproduces the state, only traveling in the opposite direction and with a change of phase for ψ_2 . The states ψ_1 and ψ_2 in this way are their own antiparticles. These self-antiparticle states are plane polarized in perpendicular directions. For

the states with momenta in the positive z direction, as given by Eq. (11) and (12), the fields are seen to be

$$E_{1x} = \cos[p\hbar^{-1}(z - ct)] \quad (19a)$$

$$B_{1y} = \cos[p\hbar^{-1}(z - ct)] \quad (19b)$$

$$E_{2y} = \sin[p\hbar^{-1}(z - ct)] \quad (19c)$$

$$B_{2x} = -\sin[p\hbar^{-1}(z - ct)] \quad (19d)$$

with all other components zero.

The final point to be demonstrated here is that only the self-antiparticle type of photon is emitted or absorbed when a system makes a transition between states of definite parity. Consider for simplicity a spinless charged particle described by a Schrödinger wave function $\psi_m(\mathbf{x}, t)$. [The subscripts m and γ are used for the material particle and the photon.] Suppose the particle is bound in some system and makes a transition from an initial state i to a final state f , both eigenstates of parity P , with emission or absorption of a photon. As is well known the transition probability is determined by the interaction integral

$$I = -(e/Mc) \int d^3x [\psi_{mf}^*(\mathbf{x}, t) \mathbf{p} \psi_{mi}(\mathbf{x}, t)] \cdot \mathbf{A}(\mathbf{x}, t) \quad (20)$$

where e and M are the charge and mass of the particle and \mathbf{A} is the vector potential of the photon in the Coulomb gauge,

$$\nabla \cdot \mathbf{A} = 0 \quad (21)$$

Here and below, the integrals extend $\overline{\mathbf{r}}$ over all space. The fields are found from the potential by the relations

$$\mathbf{E} = -c^{-1} \partial \mathbf{A} / \partial t \quad (22)$$

$$\mathbf{B} = \nabla \times \mathbf{A} \quad (23)$$

To make the argument, one first expresses the interaction explicitly in terms of the fields. The potential is found from the fields by integrating this way:

$$\mathbf{A}(\mathbf{x}, t) = \frac{1}{4\pi} \nabla \times \int d^3y \frac{\mathbf{B}(\mathbf{y}, t)}{|\mathbf{x} - \mathbf{y}|} \quad (24)$$

It is easily verified that this expression for \mathbf{A} satisfies Eqs. (21), (22), and (23) by using Eqs. (1), (3), and the fact that $\nabla^2 |\mathbf{x} - \mathbf{y}|^{-1} = -4\pi \delta(\mathbf{x} - \mathbf{y})$. Also it is assumed that fields of interest will be zero outside a finite region of space so that in making partial integrations there are no contributions from infinity. Then by using Eq. (24) and replacing \mathbf{B} by $(\psi_\gamma - \psi_\gamma^*)/2i$, one can rewrite the interaction integral as

$$I = \frac{ie}{8\pi Mc} \int d^3x [\psi_{mf}^*(\mathbf{x}, t) \mathbf{p} \psi_{mi}(\mathbf{x}, t)] \cdot \nabla \times \int \frac{d^3y}{|\mathbf{x} - \mathbf{y}|} [\psi_\gamma(\mathbf{y}, t) - \psi_\gamma^*(\mathbf{y}, t)]$$

However, if i and f are eigenstates of parity, then,

by changing integration variables from \mathbf{x} and \mathbf{y} to \mathbf{x} and $-\mathbf{y}$ in the ψ_i^* term, one sees that

$$I = \frac{ie}{8\pi Mc} \int d^3x [\psi_m^*(\mathbf{x}, t) \mathbf{p} \psi_m(\mathbf{x}, t)] \\ \cdot \nabla \times \int \frac{d^3y}{|\mathbf{x} - \mathbf{y}|} (1 + KP) \psi_i(\mathbf{y}, t)$$

where the factor is $(1 - KP)$ if i and f have the same parity, $(1 + KP)$ if i and f have opposite parity. Since

$$KP(1 \mp KP) \psi_i = \mp (1 \mp KP) \psi_i,$$

only the type of photon that is its own antiparticle can be involved in the transition in either case.

R. H. GOOD, JR.

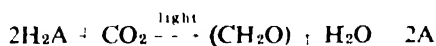
References

- Heitler, W., "The Quantum Theory of Radiation," London, Oxford University Press, 1954. Heitler discusses the properties of photons from different points of view than used here and especially shows various techniques for the quantization of the electromagnetic field.
- Good, R. H., Jr., *Rev. Mod. Phys.*, **28**, 659 (1960). A non-mathematical pictorial discussion of the different types of photons is given.
- Good, R. H., Jr., in Brittin, W. I., and Dunham, J. G., Eds., "Lectures in Theoretical Physics," New York, Interscience Publishers, 1959. A more complete treatment of the subject is given, and the theories of other massless particles are also treated.

Cross-references: BOSE-EINSTEIN STATISTICS AND BOSONS, ELECTROMAGNETIC THEORY, LIGHT, MATRICES

PHOTOSYNTHESIS*

Photosynthesis is briefly summarized and slightly oversimplified by the general equation



where H_2A represents a hydrogen donor and (CH_2O) represents carbohydrate. In green plants, H_2A is water and A is oxygen; in photosynthetic bacteria, H_2A may be any number of reducing substances (e.g., H_2S and ethanol) but is *never* water. The ability of all photosynthetic organisms (both plants and bacteria) to reduce CO_2 to carbohydrate is derived from the ability to utilize the energy of an absorbed photon in a photochemical reaction which produces a powerful reductant X_1 (see Fig. 1) capable of driving the enzymic reactions involved in CO_2 fixation. The ability of green plants (including algae) to evolve oxygen from water is derived from their ability to carry out a second photochemical reaction which produces an oxidant Y_2 (see Figure) capable of

oxidizing water to oxygen. Photosynthetic bacteria lack this second photochemical system.

The primary action of light in photosynthesis is the excitation of pigment molecules. The energy of the photon is first converted to electronic excitation energy, and subsequently stored as chemical energy. Since the number of pigment molecules in general far outnumber the number of primary reactant molecules, the efficient conversion of electronic excitation energy to chemical energy requires efficient transfer of excitation from the site of photon absorption to the site of primary photochemical action. The process generally favored in accounting for this efficient energy transfer is weak resonance interaction, and the array of pigment molecules which feed excitation energy to a single reaction center is known as a photosynthetic unit. Resonance transfer is a nonradiative process which depends upon the overlap of the fluorescence emission spectrum of the donor molecule and the absorption spectrum of the acceptor molecule as well as the separation distance. In the photosynthetic unit, singlet-state excitation resulting from photon absorption migrates from molecule to molecule until trapped by localization in a molecule whose lowest singlet state is below the lowest singlet state of any of its nearest neighbors. This is the reaction center. At the reaction center are the primary reactant molecules which are involved in the primary photochemical oxidation-reduction reaction, the nature of which still remains a matter of lively experiment and debate.

The pigments universally associated with photosynthesis are the chlorophylls. Chlorophyll *a* appears to be essential for oxygen-evolving photosynthesis, and its close relative, bacteriochlorophyll, seems to be required for all bacterial photosyntheses. Other types of chlorophyll as well as nonchlorophyllous pigments have been shown to function as accessory pigments which absorb light and transfer excitation energy to chlorophyll *a* or bacteriochlorophyll. As an example of how the various pigments function, the case of the green photosynthetic bacteria will be cited.⁶ The predominant pigment which serves as the major light absorber in green bacteria is chlorobium chlorophyll with broad absorption bands at wavelengths of 440 and 750 nanometers (nm). A lesser component is bacteriochlorophyll with bands at 371 nm, 603 nm and 809 nm. The reaction center, present in very low concentration, has a band at 840 nm. A quantum of blue light ($\lambda = 440 \text{ nm}$, $E = 2.8 \text{ eV}$) absorbed by one of 200 chlorobium chlorophyll molecules is immediately dropped in energy to 1.7 eV corresponding to the lowest excited singlet state. This excitation energy may be transferred several times between adjacent chlorobium chlorophyll molecules until eventually the transfer to bacteriochlorophyll at 1.5 eV confines the excitation to an array of approximately 20 molecules of which one is the reaction center of energy 1.4 eV. The trapped excitation energy is then converted to chemical energy ($\sim 1 \text{ eV}$) and heat.

In both algae and photosynthetic bacteria,

* Support by the U.S. Atomic Energy Commission is acknowledged.

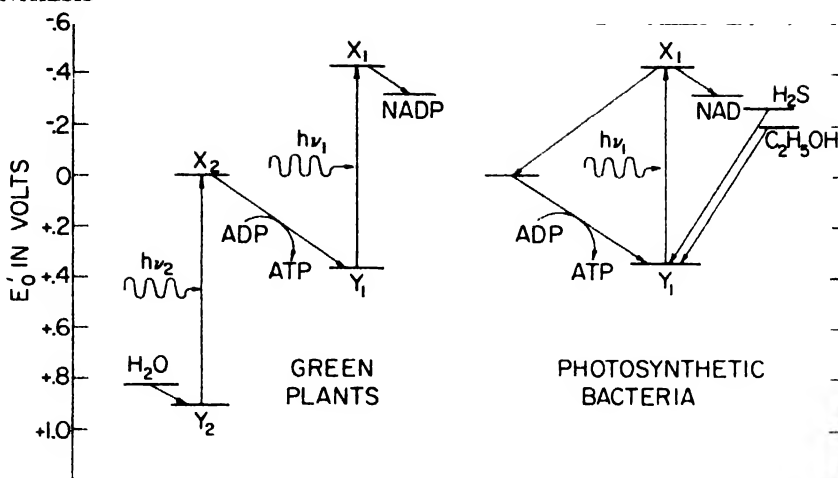
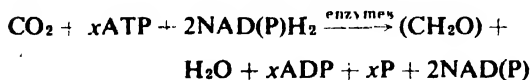


FIG. 1.

most of the measured quantum requirements for photosynthesis have been in the range of 9 to 12 quanta per CO_2 . In algae, the theoretical requirement based on two photochemical reactions in series is 8 quanta per CO_2 or 2 quanta per reducing equivalent. At first glance it would seem that the quantum requirement for bacterial photosynthesis should be half of that for oxygen-evolving photosynthesis, since the bacterial photosynthesis requires only one photochemical reaction. The experimental evidence, however, favors the view that the quantum requirement is about the same in each case. One possible explanation is outlined in the Figure in which the straight arrows denote the transfer of electrons (or equivalent hydrogen atoms) in oxidation-reduction reactions. The curved arrows represent the storage of chemical energy in phosphorylation reactions coupled to electron transfer.

This explanation is based on the well-documented and generally accepted concept that CO_2 reduction in photosynthesis results from a series of enzymic reactions which require reducing power in the form of reduced nicotinamide-adenine dinucleotide (NADH_2) or its phosphate (NADPH_2) and additional chemical energy in the form of adenosine triphosphate (ATP). The photochemical reactions of photosynthesis need only produce sufficient NADH_2 (or NADPH_2) and ATP to drive the following over-all reaction:



According to the hypothesis shown in Fig. 1, the light-driven production of ATP in bacteria competes with the production of reducing power, whereas in oxygen-evolving photosynthesis, ATP production is coupled to the exergonic reactions which link the reductant X_2 , produced by photochemical reaction 2, to the oxidant Y_1 , produced by photochemical reaction 1.

JOHN M. OLSON

References

1. Rabinowitch, E. I., "Photosynthesis and Related Processes," New York, Interscience Publishers, 2088 pp. Vol. I, "Chemistry of Photosynthesis, Chemosynthesis and Related Processes *in Vitro* and *in Vivo*," 1945; Vol. II, Part I, "Spectroscopy and Fluorescence of Photosynthetic Pigments; Kinetics of Photosynthesis," 1951; Vol. II, Part 2, "Kinetics of Photosynthesis" (cont'd); Addenda to Vol. I and Vol. II, Part I, 1956.
2. Duysens, L. N. M., "Energy Transformations in Photosynthesis," *Ann. Rev. Plant Physiol.*, 7, 25-50 (1956).
3. Clayton, R. K., "Photosynthesis: Primary Physical and Chemical Processes," *Ann. Rev. Plant Physiol.*, 14, 159-180 (1963).
4. Bassham, J. A., "Energy Capture and Conversion by Photosynthesis," *J. Theoret. Biol.*, 4, 52-72 (1963).
5. Kamen, M. D., "Primary Processes in Photosynthesis," New York, Academic Press, 1963, 183 pp.
6. Gest, H., San Pietro, A., and Vernon, L. P., Eds., "Bacterial Photosynthesis," Yellow Springs, Ohio, Antioch Press, 1963, 523 pp.
7. "Photosynthetic Mechanisms of Green Plants," Publ. 1145, NAS-NRC, Washington, D. C., 1963, 766 pp.
8. Bay, Z., and Pearlstein, R. M., "A Theory of Energy Transfer in the Photosynthetic Unit," *Proc. Natl. Acad. Sci. (U.S.)*, 50, 1071-1077 (1963).

Cross-references: LIGHT, ULTRAVIOLET RADIATION.

PHOTOVOLTAIC EFFECT

The photovoltaic effect (PVE) is the generation of an emf as a result of the absorption of light. Three phenomena are involved in the effect. The first of these is photoionization, i.e., the generation of equal numbers of positive and negative charges by light absorption. The second is the migration of one or both of the photo-liberated charges to a region where separation of the positive and negative charges can occur.

The third is the presence of a charge-separation mechanism. The photovoltaic effect can occur in gases, liquids and solids, but it has been studied most intensively in solids and, therefore, this discussion will be limited to solids, especially semiconductors.

Photoionization. Semiconductors and insulators are characterized by a threshold energy for photoionization, which is identical with the forbidden energy gap of the material E_g . Thus, photoionization can occur only if the light contains photons whose energy exceeds E_g . Values of E_g range from several electron volts (which would require ultraviolet photons for ionization) to small fractions of an electron volt (infrared photons could cause ionization). Photoionization in metals can also lead to a photovoltaic effect. In this case, the photoionization occurs near the surface and the photo-liberated carriers escape over the barrier at the metal surface (whose magnitude in electron volts will be denoted by W) and enter an abutting semiconductor or insulator, or alternatively, the electrons can be emitted into a vacuum surrounding the metal.

Now the absorption of light in any substance follows Lambert's law which states that

$$N(x) = N(0) e^{-\alpha x}$$

where $N(0)$ and $N(x)$ are the numbers of photons per unit area at a reference point (0) and at a distance x along the direction of propagation of the light beam and α is the absorption constant. Figure 1 shows how α changes with photon energy in a number of semiconductors used in photovoltaic cells. This dependence of α on photon energy (i.e., the color of the light) is of course intimately related to the dependence of the photovoltaic effect on photon energy.

Migration of the Photo-liberated Charges. Because of Lambert's law, the photo-liberated

charges are distributed along the path of the light beam. Normally, they would move about at random until they recombine with carriers of opposite sign. This recombination process is characterized by a mean free lifetime of a pair, which must be large enough to permit the carriers to move to a charge-separation region since the magnitude of the photovoltaic effect depends on the number of charges which move to the charge separation region. The photo-liberated carriers move either by diffusion or under the action of a built-in electrostatic field.

Charge-separation Mechanism. Charge separation requires a change in electrostatic potential between two regions of the solid, so that when a pair of opposite charges migrate to the region of the potential change, one of them can lower its potential by moving across this region. A large photovoltaic effect requires that the change in potential should be large and that it should occur over a distance which is short compared to the mean distance a free carrier can travel before recombination. These requirements imply the presence of a dipole layer which, in turn, implies an abrupt change in some property of the material. Such a dipole layer can occur either at the surface or in the interior of the material. A surface barrier usually involves a metal-semiconductor contact, a contact between the semiconductor and its oxide or, more generally, a contact with some other semiconductor. A barrier inside a material implies an abrupt change in the conductivity, which in the extreme case involves a change of conductivity type, as in a $p-n$ junction.

Photovoltaic Effect at $p-n$ Junctions. The remainder of this article is devoted to the specific case of the photovoltaic effect at a $p-n$ junction. A $p-n$ junction is formed by arranging the chemical impurity distribution in a single crystal of a semiconductor so that electric current is carried primarily by electrons on one side of the junction (the n -side) and primarily by holes on the other side (the p -side). The resulting electrostatic potential profile is such that excess holes can lower their energy by moving from the n - to the p -side while excess electrons lower their energy by moving in the opposite direction. Light absorption in either region leads to an increase in the concentrations of both holes and electrons, but the $p-n$ junction limits the flow of carriers of a given sign to a single direction, and therefore, separates the electrons and holes. If a resistive load is connected between the p - and n -regions, the current through the load I_L is represented adequately by the equation

$$I_L = I_s - I_0(e^{\Delta V} - 1)$$

where I_0 and Δ are functions of the absolute temperature, material parameters and the fundamental constants; I_s is the photo-generated current which would flow if the junction were short circuited. The parameter I_s is a function of the absorption constant α , the carrier lifetime τ , the spectral composition of the light and the geometry of the junction.

Applications. The photovoltaic effect can be

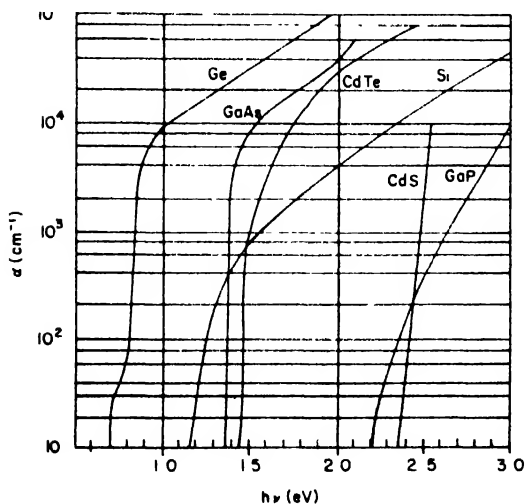


FIG. 1.

used to convert light to electricity and, indeed, p - n junction photovoltaic cells in silicon have been the principal power sources on space satellites. The effect can of course be used to detect small amounts of radiation, and its sensitivity compares favorably with that of photoconductive detectors. The ionizing radiation need not be light so that this effect can also be used to detect x-rays, fast electrons, protons, etc. Studies of the spectral response of the photovoltaic effect can yield valuable information about basic material parameters like α , τ , E_g , σ , etc. The photovoltaic effect provides a convenient tool for studying radiation damage in semiconductors. It is also a basic phenomenon in phototransistors.

JOSEPH LOFERSKI

Reference

- Marton, L. B., Ed., "Methods of Experimental Physics," Vol. 6B, New York, Academic Press, 1959.
 Angrist, S. W., "Direct Energy Conversion," Boston, Allyn and Bacon, 1965.

Cross-references: ENERGY LEVELS, PHOTOCONDUCTIVITY, PHOTOELECTRICITY, SEMICONDUCTORS.

PHYSICAL ACOUSTICS

Sound waves in the ideal sense are mechanical disturbances propagating in a continuous medium. The basic fact of interest from a physical standpoint is that these waves carry momentum and energy without a net transport of mass. However, there are two other facts worth mentioning which are that sound waves propagate in almost all substances and that sound waves can be used to study the physical properties of, or to cause physical changes in the substances through which they propagate.

In practice, the approximation that natural materials are continuous fails when the wavelength is short enough to be comparable to interatomic or intermolecular distances--in a gas, the mean-free-path of a molecule; in a solid, the distance between atoms. The region of validity of the approximation that the waves are of small amplitude is more difficult to circumscribe since it depends on more than one property of the medium. In most cases, the approximation is valid when the disturbance in the velocity of the medium as the wave passes is small compared to the speed of sound in the medium. In air under normal conditions, this will occur when the pressure disturbance becomes an appreciable fraction of atmospheric pressure. In liquids and solids, however, pressure disturbances of hundreds or thousands of atmospheres propagate as small-amplitude waves. The low compressibility, or high "stiffness," of these media increase the speed of sound and at the same time require a larger pressure fluctuation to produce a given velocity fluctuation. If the medium is in a critical

state, however, the definition of what constitutes a small disturbance can depend on other factors. For example, cavitation in a liquid occurs at particular sites that are called cavitation nuclei. While the reasons for the existence of these sites in a fluid are not completely understood, it is clear that they are regions that are subject to fracture at pressures far below those that would fracture the pure liquid.

The energy transported by the wave per unit area normal to the direction of propagation is called the intensity and can be measured in watts per square meter (the mixed unit, watt per square centimeter, is also in common use). A sound wave in air of intensity comparable to that of sunlight would be just beyond the pain threshold of the ear, so the energy of sound waves is not normally felt. The principle of conservation of energy applied to the energy transported by the waves, however, is a powerful investigational tool. The momentum transport per second per unit area is the intensity divided by the wave speed. Since this is also true for electromagnetic waves, it follows that a sound wave in air (speed 330 m/sec) of a given intensity transports about a million times more momentum than an electromagnetic wave of the same intensity. When the wave is absorbed or reflected, there is a radiation pressure on the absorber or reflector equal to the rate of change of momentum per unit area that is I/c on a perfect absorber and $2I/c$ on a perfect reflector, where I is the intensity and c the speed of sound. The measurement of this pressure is a fundamental method for measuring the amplitude of the wave.

The basic equations for a continuous medium are the continuity equation, which is an expression of the law of conservation of mass, and the Navier-Stokes equation, which relates the time rate of change of momentum in each volume element to the forces on the element. An additional equation is required that relates the scalar pressure forces to the other variables. While this third equation must be derived from the laws of thermodynamics, a relationship that is widely applicable is

$$\rho = \kappa \rho_0 p$$

where κ is the compressibility of the medium, ρ_0 is the ambient density, and p and ρ are the disturbances in pressure and density associated with the wave. This relationship is useful as long as the wave does not change the value of κ appreciably. If there is no dissipation in the medium or if the dissipation is viscous, p and ρ will be in phase. In cases where dissipation exists due to irreversible heat flow, or the stress-strain relationship contains time derivatives, the equation can be retained by allowing κ to become complex. There are then three equations in three dependent variables which are p , ρ and the velocity disturbance u .

The equations have a unique solution only if certain boundary conditions are applied. In the case of an unbounded medium, the boundary condition is that all sources of sound must

radiate energy outward, and this is sufficient to produce a unique solution. In a bounded region, reflections must be taken into account. A sufficient boundary condition is that the impedance of the boundary, defined as the ratio of the pressure disturbance to the normal component of the velocity disturbance, be specified at each point on the boundary. In simple cases, the boundary conditions can be satisfied by a superposition of elementary solutions. The impedance tube is an example of a case where plane waves are used to measure the reflective and absorptive properties of a sample placed at one end of the tube in terms of the incident and reflected waves. The details of the distribution of sound in an auditorium, however, can be a complex problem. Another approach is to reformulate the equations as an integral equation containing the boundary conditions explicitly. This integral equation, a mathematical expression of Huygens' principle, is known in acoustics as Kirchhoff's formula. A third approach is to include the boundary conditions in the differential equations. An example of this procedure is the propagation of sound in a duct of varying cross section. If the change in width of the duct per wavelength is small compared to the wavelength of sound in the duct, the shape of the duct can be introduced into the continuity equation. This procedure leads to the "horn equation" which describes sound propagation in flared ducts. The subject of boundary conditions is further complicated by the fact that waves may propagate in the boundary itself. A boundary characterized by the fact that each point behaves independently is called a locally reacting boundary. If neighboring parts of a boundary interact to support wave motion, the boundary is non-locally interacting, and if the wave speed in such a boundary is higher than in the medium, energy can propagate along the boundary and leak back into the medium. In such a case, the wave may find a shorter time travel path through the boundary than through the medium itself.

Sound waves are generated in regions where time-varying forces act on the medium. While in some cases it may be convenient to think in terms of boundary conditions, this is not always convenient or even possible. A moving piston, such as the cone of a loudspeaker, can be thought of as a boundary condition in that the normal velocity, or pressure gradient, is specified. Sound generated by a region of turbulence, however, cannot be treated in this fashion because the source strength is distributed over the entire turbulent volume and not restricted to a surface. In fact, in the region of turbulence it is difficult to separate the motion of the medium into hydrodynamic and acoustical components. If the mechanism of sound generation is confined to a definite region, it is natural to describe the sound field at a distance in terms of spherical harmonics, and the amplitudes of these harmonics can be related to integrals of moments of the motion of the medium taken over the source region. The most simple natural source, the

monopole, can be described as a periodic injection and removal of mass at some point in the medium. In a stationary medium, the wave fronts are concentric spheres and the pressure disturbance is proportional to the time rate of change of mass flow. The next higher spherical harmonic, the dipole, is related to the net fluctuating force on the medium. For example, if an airfoil is subject to fluctuating lift and drag forces, there are equal and opposite forces exerted on the medium. These forces give rise to dipole sound radiation, and the sound pressure is proportional to the time rate of change of the forces. The sound fields generated by turbulence and by earthquakes are related to the integral of shear motions over the source volume and have a quadrupole distribution.

If the ambient properties of the medium are everywhere the same, the medium is said to be homogeneous. A uniform translational motion of the medium, however, deserves special attention. While a suitable coordinate transformation can remove a translational motion without changing the basic equations, such a transformation does not remove relative motion between the medium and any sources or receivers, and such motion causes the medium to be anisotropic. Although a receiver is never completely passive, the usual approximation is to assume that it is. In this case, the receiver has no physical effect on the medium and simply registers what it sees. A source, on the other hand, is contributing energy and momentum to the medium, and the distribution of these quantities is affected by the motion. Surfaces of constant phase are carried along by the medium in the same way as ripples from a pebble dropped in moving water, and for a monopole source at rest in a moving medium, they form spheres centered a distance RU/c downstream from the source, where R is the radius of the sphere, U the speed of the medium, and c the speed of sound in the medium at rest. The surfaces of constant pressure for such a source, on the other hand, are ellipsoids of revolution centered at the source with the minor axis in the flow direction. It follows that the intensity is greater to one side than it is at an equal distance either upstream or downstream. In general, the problem is complicated by the fact that the energy flow vector is not perpendicular to a surface of constant phase.

The dispersion relationship for acoustic plane waves relates the phase velocity and attenuation (or growth) of these waves to physical properties of the medium.

Classically, the important parameters for a fluid are viscosity and heat conduction. While a general solution including these parameters leads to a sixth-order polynomial, in cases where they are small the problem can be solved to a good approximation by three pairs of roots representing two plane waves, two viscous shear waves, and two thermal waves. The plane waves, traveling to the left and right, are attenuated slightly by viscosity and heat conduction. The shear waves and thermal waves are not excited in the

bulk of the medium. At boundaries, the shear and thermal waves will exist, however, to satisfy tangential velocity and thermal boundary conditions. While the viscous and thermal waves exist only in thin boundary layers (the real and imaginary parts of the propagation constant are equal), they may be responsible for a major part of the absorption. For example, in the case of plane waves propagating in air in a pipe the ratio of boundary absorption coefficient to bulk absorption coefficient is about $10^8/Lf^{3/2}$ where L is the ratio of the area of the pipe to its perimeter and f is the frequency. For very narrow pipes, the ratio is larger.

In general, the measured value of attenuation of plane waves exceeds the classical predictions based on heat conductivity and viscosity. The discrepancy varies not only from one substance to another but also may depend on the past history of the substance. The additional attenuation is caused by (and reflects) interactions between particles on a small scale compared to a wavelength—in other words, between individual atoms, molecules, or groups of molecules. The problems are as various as the chemical properties of matter.

In solids, it has been shown that dissipation usually is related to departures from ideal crystalline structure on a relatively large scale. Annealing greatly reduces the attenuation of sound. In fluids, it has been shown, however, that attenuation can be related to interatomic binding forces as well as the forces binding clusters of molecules. The vibrational relaxation of O_2 in air and of magnesium salts in water, for example, can account for the excess attenuation of sound waves in the atmosphere and in seawater.

L. WALLACE DEAN, III

References

- Morse, P. M., "Vibration and Sound," Second edition, New York, McGraw-Hill Book Co., 1948.
 Rayleigh, "Theory of Sound," New York, Dover, 1945.
 Blokhintzev, D., "Acoustics of an Inhomogeneous, Moving Medium" NACA TM1399 (1956).
 Lighthill, M. J., "On Sound Generated Aerodynamically," *Proc. Roy. London Ser. Soc. A*, **211**, 564 (1952).
 Mason, W. P., Ed., "Physical Acoustics," New York, Academic Press (planned in seven volumes not all of which have been published at this date).

Cross-references: ACOUSTICS; ARCHITECTURAL ACOUSTICS; CAVITATION; ELECTROACOUSTICS; MUSICAL SOUND; RESONANCE; ULTRASONICS.

PHYSICAL CHEMISTRY

Physical chemistry emerged as a separate discipline in the second half of the nineteenth century, van't Hoff, Ostwald and Arrhenius contributed conspicuously in laying its foundation; van't Hoff is especially remembered for

his explanation of osmotic pressure. According to him, a solute in a vessel separated from pure solvent by a membrane permeable only to the solvent exerts the same pressure which the same number of moles of gas would if confined in the same volume. In exerting this pressure, the solute draws solvent through the semipermeable membrane except as it is restrained by a pressure equal to the osmotic pressure. He also codified equilibrium theory in his celebrated "Études de Dynamique Chimique" published in 1884. It is interesting that van't Hoff received the first Nobel prize for chemistry in 1901.

Ostwald's dilution law and Arrhenius' theories of reaction rates and of the ionization of electrolytic solutions played key roles in getting physical chemistry off to an impressive beginning. Willard Gibbs' development of thermodynamics and statistical mechanics and van der Waals' equation of state are further milestones in the onward march of the new discipline.

Physical chemistry is the attempt to codify into laws the many qualitative observations made on all types of molecular systems. Such an ambitious goal can only be achieved by using the full spectrum of experimental and theoretical procedures. Atoms are known to consist of positive nuclei surrounded by electrons circulating in orbits about the nuclei. Since the electrostatic and magnetic laws of force are known for the atoms and since with wave mechanics we can express the laws of motion of particles in terms of the known potentials, it follows that the laws of chemistry are known. As a consequence of this, all properties of atoms and molecules can be calculated exactly, in principle. Nevertheless, such calculations are so difficult that only for the simplest systems can the calculations be carried out exactly. However, even for complicated systems, wave mechanics provides the framework into which all observations can be fitted.

Bohr's Model of the Atom. It seems useful to trace quantum mechanics briefly from its beginnings. In 1913, Bohr working in Rutherford's laboratory, postulated that the angular momentum, $mr^2\omega$, of an electron moving in a circle about a nucleus, when multiplied by the number of radians, 2π , in the circular orbit, must equal some integer n times Planck's unit of action h . Accordingly he wrote

$$mr^2\omega 2\pi = nh \quad (1)$$

He also equated the centrifugal force, $mr\omega^2$, to the coulomb force of attraction, e^2/r , to give

$$mr\omega^2 = \frac{e^2}{r^2} \quad (2)$$

Finally the total energy E is the sum of the kinetic energy $\frac{1}{2}mr^2\omega^2$ and the potential energy $-e^2/r$, that is

$$E = \frac{1}{2}mr^2\omega^2 - \frac{e^2}{r} = -\frac{1}{2}\frac{e^2}{r} \quad (3)$$

Here, m , r and ω are the reduced mass of the electron, the radius of the electron orbit, and the angular momentum, respectively. Solving these equations we find for the radius

$$r = \frac{n^2 h^2}{4\pi^2 m e^2} \quad (4)$$

and for the energy

$$E = - \frac{2\pi^2 m e^4}{n^2 h^2} \quad (5)$$

The number n is the principle quantum number. Equation (5) may be used to predict the spectra of hydrogen exactly, and thus constituted a great step forward in explaining atomic structure. The energy for a single electron circulating about a nucleus of charge ze is

$$E = - \frac{2\pi^2 m z^2 e^4}{n^2 h^2} \quad (6)$$

If an electron is circulating about a nucleus whose charge is ze , but whose effective charge is reduced to $(z-s)e$ by a screening cloud of other electrons, there is a corresponding reduction of the electron's binding energy to

$$E = - \frac{2\pi^2 m e^4 (z-s)^2}{n^2 h^2} \quad (7)$$

Here, s is called the screening constant. Equation (7) explains much of chemistry. In spite of this agreement, Bohr's theory is incomplete since it provides no procedure for calculating the screening constant.

The De Broglie Wavelength. In the middle twenties, this complication was resolved by the advent of quantum mechanics. The matrix theory of Heisenberg was followed by De Broglie's interpretation of particle motion as wave motion. In the new theory, Bohr's quantization of orbits as formulated in Eq. (1) was replaced by the idea that the circumference of a circular orbit, $2\pi r$, must be an integral number of wavelengths, $n\lambda$. Thus,

$$n\lambda = 2\pi r = \frac{n^2 h^2}{2\pi m e^2} \quad (8)$$

Substituting into Eq. (8) the value for the momentum, p , obtained from the kinetic energy equation $p^2/2m = \frac{1}{2}e^2/r$ leads to the famous De Broglie relation

$$\lambda = h/p \quad (9)$$

This same result applies to photons, which have both a particle and a wave aspect. Thus, according to Einstein

$$E = h\nu = mc^2 = mc\lambda\nu \quad (10)$$

Whence

$$\lambda = \frac{h}{mc} = h/p \quad (11)$$

Confirmation of the De Broglie wavelength for electrons was soon provided by Davisson and Germer and by G. P. Thomson who demonstrated that electrons were in fact diffracted by crystalline materials much as x-rays are.

The Wave Equation. Schrödinger took the next step by substituting into the wave equation

$$\frac{1}{a^2} \frac{\partial^2 \psi}{\partial t^2} = \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} \quad (12)$$

values for a , the wave velocity. Thus

$$a = \lambda\nu = \frac{hE}{p} = \frac{E}{\sqrt{2m(E-V)}} \quad (13)$$

Here ψ is the amplitude of the wave, while t is the time and x , y and z are the coordinates of position for the particle. For stationary waves, the amplitude may be expressed as

$$\psi = \phi(x, y, z) e^{-\frac{2\pi i Et}{h}} \quad (14)$$

which when substituted along with Eq. (13) into Eq. (12) leads to Schrödinger's famous time-independent equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} + \frac{8\pi^2 m}{h^2} (E - V) \phi = 0 \quad (15)$$

From Eq. (14), we note that

$$-\frac{h}{2\pi i} \frac{\partial}{\partial t} \psi = E\psi \quad (16)$$

Using this result we see that Eq. (15) may be generalized to the time dependent form

$$\left[-\frac{h^2}{8\pi^2 m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + V \right] \psi = -\frac{h}{2\pi i} \frac{\partial \psi}{\partial t} \quad (17)$$

Now since ϕ in Eq. (15) is the amplitude of the wave, the square of the amplitude $|\phi|^2$ is necessarily interpreted as proportional to the probability of a particle being at a point in space, just as the square of the amplitude of the wave expresses the density of radiation at a point in physical optics. Restricting the allowed solutions of Eq. (15) to those eigenfunctions ϕ which have acceptable values leads to the allowed values of the energy E which Bohr found for hydrogen.

It is important to note that we now have a theory that can be generalized to many particles. The operator in square brackets on the left of Eq. (17) is just the expression for the energy of a particle providing $h/2\pi i \partial/\partial x$, $h/2\pi i \partial/\partial y$ and $h/2\pi i \partial/\partial z$ are replaced by the corresponding momenta p_x , p_y and p_z respectively. Analogously $-h/2\pi i \partial/\partial t$ is the operator for the energy E or H as was noted in arriving at Schrödinger's time-dependent equation. It thus seems natural to generalize Eq. (17) to many particles by having V express the total potential energy for the system

and taking as the operator for the kinetic energy, the sum over all particles

$$\sum_i \left[-\frac{h^2}{8\pi^2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \right]$$

This is not the place to elaborate quantum mechanics in detail, but clearly it completely transforms physical chemistry from an almost exclusively experimental discipline to one for which there is a fundamental theory as well.

Mean Values of Properties. According to quantum mechanics, the mean value of any property, x , of a system in the state ψ_i having the energy ϵ_i is

$$\bar{x}_i = \int \psi_i^* x \psi_i dT$$

Here x is the quantum mechanical operator for the property x , it is obtained from the latter by replacing each momentum p_i by $h/2\pi i \partial/\partial q_i$, where q_i is the corresponding positional coordinate, and keeping the q 's and t unchanged. If there are ambiguities, x must be chosen to give real values for the calculated property. Now since the probability, p_i , of a system at equilibrium being in the state ψ_i is

$$p_i = \frac{e^{-\epsilon_i/kT}}{\sum_i e^{-\epsilon_i/kT}} \quad (18)$$

the mean value of the property, x , for the system is

$$\langle x \rangle = \frac{\sum_i \bar{x}_i e^{\epsilon_i/kT}}{\sum_i e^{-\epsilon_i/kT}} \quad (19)$$

Thus, in principle, the values of all properties can be calculated. Actually such calculations are quite difficult and are rarely carried through exactly. The sum $\sum_i e^{-\epsilon_i/kT}$ over all the states is called a partition function and is frequently represented by the symbol f .

Statistical Mechanics. Statistical mechanics whose beginnings go back to Maxwell and Boltzmann was greatly developed and systematized by J. Willard Gibbs. It provides still another cornerstone of theoretical physical chemistry.

The probability, P_i , of a system being in the state i with energy ϵ_i is already given in Eq. (18). The average energy E is accordingly

$$E = \frac{\sum_i \epsilon_i e^{-\epsilon_i/kT}}{\sum_i e^{-\epsilon_i/kT}} \equiv kT^2 \frac{d \ln}{dT} \left(\sum_i e^{-\epsilon_i/kT} \right) \quad (20)$$

But, from thermodynamics, we have for the average energy of a system

$$E = -T^2 \frac{\partial(A/T)}{\partial T} \quad (21)$$

Here, A is the Helmholtz free energy and $A_1 - A_2$ is the maximum work a system can be made to perform at constant temperature in passing from

state 1 to state 2. Equating the two values of E , given in Eqs. (20) and (21) and integrating gives

$$A = kT \ln \sum_i e^{-\epsilon_i/kT} + cT \quad (22)$$

Here c is an integration constant. Now the lowest state of the system must be nondegenerate and c must be zero in order that Eq. (22) will yield zero entropy for the system at the absolute zero of temperature as is required by the third law of thermodynamics. Thus, we can write the exciting result

$$A = -kT \ln \sum_i e^{-\epsilon_i/kT} \quad (23)$$

as the fundamental equation relating statistical mechanics and thermodynamics. If the energy levels ϵ_i are known as functions of the volume of the system, then A is known as a function of V and T , and all the thermodynamic properties of the system can be calculated.

Equilibrium Theory. The probability, P_i , of a system which obeys Boltzmann statistics, being in the i th state is equal to n_i/N where n_i is the number of systems in the i th state and N is the total number of systems. Thus we can rewrite Eq. (19) in the form

$$\frac{n_i}{e^{-\epsilon_i/kT}} = \frac{N}{\sum_i e^{-\epsilon_i/kT}} = \lambda = e^{\mu/kT} \quad (24)$$

Here λ is called the absolute activity and μ is the chemical potential. For a system at equilibrium we can write

$$aA + bB + \cdots = dD + eE + \cdots \quad (25)$$

Here, a molecules of substance A reacts with b molecules of B to give d molecules of D plus e molecules of E, etc. According to thermodynamics, a system at equilibrium is incapable of doing work so that the reaction involves no change of chemical potential.

Thus,

$$a\mu_A + b\mu_B + \cdots = d\mu_D + e\mu_E + \cdots \quad (26)$$

and

$$\lambda_A^a \lambda_B^b \cdots = \lambda_D^d \lambda_E^e \cdots \quad (27)$$

Whence

$$\frac{(n_A)^a}{f_A^a} \frac{(n_B)^b}{f_B^b} \cdots = \frac{(n_D)^d}{f_D^d} \frac{(n_E)^e}{f_E^e} \cdots \quad (28)$$

Here n_A is the number of molecules of the A th kind distributed over the partition function f_A with analogous meanings for the other symbols. If we define c_A as the concentration of A then we have

$$n_A/V = c_A \quad (29)$$

and $f_A/V = F_A$. Here F_A is the partition function per unit volume. Equation (29) can now be rewritten in the form of an equilibrium constant K_c . Thus

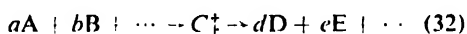
$$K_c = \frac{(C_D)^d (C_E)^e \cdots}{(C_A)^a (C_B)^b \cdots} = \frac{F_D^d F_E^e \cdots}{F_A^a F_B^b \cdots} \quad (30)$$

Now even when systems do not obey Boltzmann statistics, we still have Eqs. (26) and (27) remaining true with the auxiliary equation

$$\left(\frac{\partial A}{\partial n_A}\right)_{V, T, n_n} = \mu_A \quad (31)$$

as well as the other analogous equations. Thus equilibria for systems obeying all types of statistics are readily formulated using statistical mechanics. We now turn to rates of reaction.

Reaction Rates. A chemical reaction ordinarily proceeds slowly because it involves passage through a transition state, which is difficult to attain. The transition state lies at the saddle point of a landscape separating two valleys, one corresponding to reactants and the other to products. Such landscapes are ordinarily multi-dimensional. The path through the saddle point is called the reaction coordinate. A system at the saddle point is called the activated complex and is like other molecules except for its fleeting passage along the reaction coordinate. The reaction coordinate of the activated complex can be treated as a translational degree of freedom if due account is taken of reflection and barrier leakage. Since the transition state is the point of no return, it is the logical region for calculating the rate of reaction. If we rewrite Eq. (25) to include the activated complex C^\ddagger , we have



The rate of reaction in the forward direction, r_f , at equilibrium is

$$r_f = \mathcal{H} \frac{1}{2} C_{\delta^\ddagger} \bar{u} \cong \mathcal{H} \frac{1}{2} C^\ddagger \frac{(2\pi mkT)^{1/2}}{h} \delta \left(\frac{2kT}{\pi m}\right)^{1/2} \frac{1}{\delta} \\ = \mathcal{H} C^\ddagger \frac{kT}{h} = \mathcal{H} F^\ddagger \frac{kT}{h} \lambda_r^\ddagger \quad (33)$$

here \mathcal{H} is the transmission coefficient which corrects the rate for reflection, barrier leakage and non-equilibrium; \bar{u} , the average velocity across the barrier, has the value $(2kT/\pi m)^{1/2}$; C_{δ^\ddagger} is the concentration of activated complex per length δ along the reaction coordinate while C^\ddagger is the same quantity per single quantum state along the reaction coordinate. F^\ddagger and λ_r^\ddagger are the partition functions per unit volume and the absolute activity of the activated complexes moving in the forward direction, respectively. The net rate is thus

$$r = \mathcal{H} \frac{kT}{h} F^\ddagger (\lambda_r^\ddagger - \lambda_b^\ddagger) \quad (34)$$

Here λ_b^\ddagger is the absolute activity of reactants moving in the backward direction and may be replaced by the product of the absolute activities of the products of reaction while λ_r^\ddagger may be replaced by the products of the absolute activities of the reactants; r is of course zero at equilibrium.

Experimental Methods. In recent times, physical

chemistry has added many new techniques to the standard procedures which were used earlier in such fields as solution chemistry and in gas kinetics. Analysis of the products of reaction by mass spectrography has solved many difficult kinetic problems. Computer techniques have greatly speeded up structure determination by x-rays and by other diffraction techniques using electrons and neutrons. The development of radar in wartime provided a basis for microwave spectroscopy. Microwaves with wave lengths around one centimeter can be used to study the inversion of ammonia and internal rotations around C—C and other bonds. Microwave spectroscopy, because of the extremely long wavelengths involved, brings phenomenal precision into molecular structure measurements.

Nuclear magnetic resonance, measuring as it does the magnetic moments of atomic nuclei, reveals the different diamagnetic and paramagnetic environments surrounding atoms. For example, one can distinguish various environments of hydrogen and rates of migration between such environments. Electron spin resonance yields parallel information about the unpaired electrons. Thus free radicals are detected, and changes involving them are followed.

Chromatography makes use of the differential distribution of the components of a mixture between a stationary and a moving phase separating the components because of the different times that they require to traverse a column. At low temperatures, even *ortho*- and *para*-hydrogen can be separated by helium moving past carbon. Chromatographic separation is in fact one of the milestones in chemistry.

Valence and Molecular Structure. The electron pair bond of G. N. Lewis is explained in terms of quantum mechanics and the Pauli Exclusion Principle. Thus quantum mechanics is used to find the best bonding orbitals and the Exclusion Principle requires that only two electrons having unlike spins can occupy the orbital. The recent discovery of fluorides of the rare gases shows that novel findings are still possible in valence and molecular structure. Pauling has given us an interesting presentation of valence in his book with that title.

Thermodynamics. The thermodynamics of solutions is developing along the lines of providing mechanistic models to explain activity coefficients at the higher concentrations not explained by Debye-Hückel theory. This gives added insight into the structure of solutions. Better models are being developed to explain simple liquids. The best of them yield very realistic calculations of the properties of liquids.

Concluding Remarks. The boundaries between disciplines regularly grow more blurred with time. This is especially true of the border between chemistry and physics. Here physical chemistry or chemical physics continues to spread into new fields with no ends in sight. Physical chemistry's second century should be even more spectacular than the first has been.

HENRY EYRING

Cross-references: CHEMICAL KINETICS, CHEMICAL PHYSICS, CHEMISTRY, MICROWAVE SPECTROSCOPY, OSMOSIS, STATISTICAL MECHANICS, THERMODYNAMICS.

PHYSICS

Physics can be defined as the branch of natural science that treats those phenomena of material objects included in the subjects of mechanics, properties of matter, heat, sound, light, electricity and magnetism, and molecular and atomic processes. It describes and correlates energy and radiation, and has sometimes been defined as the study of matter and energy. While such definitions attempt to distinguish between physics and other sciences, it should be stressed that physics is not set off by itself, but is very closely related to other sciences, and there is a good deal of overlap with their spheres of interest. In many cases, the differences are merely matters of emphasis. Chemistry stresses the regrouping of atoms to form new compounds, while physics picks a particular substance and studies its behavior. Biology restricts its interest to living organisms, while physics largely omits them from its consideration. Geology confines its investigations to the earth itself and the inanimate material of which it is composed. Astronomy, on the other hand, finds its field of interest in the vast expanses of space which run on and on away from the earth.

The extent to which these sciences overlap is shown by the special branches of science dealing with borderline areas. Astrophysics treats physical phenomena as they occur in regions beyond the immediate vicinity of the earth. It includes studies of the physical processes in stars and of the nature of the radiations they emit. Physical chemistry, as its name implies, covers phenomena common to both physics and chemistry. Energy relations in chemical reactions and the effects of changes in temperature or pressure on the reactions are samples of the topics included. Geophysics relates physics and geology, and involves studies of the earth's magnetism and the effects of high pressure in the earth's interior. Biophysics deals with the physical processes that are of specific interest to the study of living organisms. Each of these fields has grown to include a vast array of scientific knowledge, and each is discussed in a separate article in this book.

The field of engineering is closely related to physics and other sciences. In general, it can be said that as knowledge about some specific topic grows to the point where use can be made of it in every-day life or in industrial processes, the topic passes over to engineering, and engineers develop and improve things using the basic information. Scientists concern themselves with the basic phenomena themselves, while engineers direct their efforts toward the solution of problems dictated by specific applications.

Historical Background. The word "physics" can be traced to the Greek word *physos*, meaning nature. All studies about nature were grouped together by the Greek philosophers, but in many

cases theories suffered from a lack of experimental research, so that the name, "natural philosophy" was more descriptive of their efforts than a name such as "natural science" would have been.

A summary of the important features in the development of physics from its early beginnings to the present advanced state can be found in the article on HISTORY OF PHYSICS.

Major Divisions of Physics. For purposes of study, physics can be reasonably well divided into major areas, which will be discussed briefly here and in greater detail in other articles in this book.

Mechanics is the science of the motions of material bodies, the forces which produce or change these motions, and the energy relations involved. Newton's three laws of motion and his law of gravitation form the foundation of this study. Work, momentum, vibration, wave motion, pressure, elasticity, and viscosity are other topics considered to be parts of mechanics.

Heat is the energy which an object possesses by virtue of the motions of the molecules of which it is composed plus the potential energy resulting from interatomic forces. The field is of great importance to other areas of learning, as well as to physics, and appears as a topic in chemistry and engineering. The term heat is also used in a different but related sense to indicate energy in the process of transfer between an object and its surroundings because a difference exists between their temperatures. Thermodynamics is the name given to the study of the relationships between heat and mechanics.

Acoustics is the science of sound. Sometimes wave motion is studied under acoustics rather than under mechanics because sound provides good illustrations of wave motion. Objects which are in vibrating motion in a medium can set that medium in motion, and the disturbance can travel through the medium in the form of a wave. Energy is transferred without the transfer of particles of the medium. Sound is the type of wave motion which has such a frequency that, if it reaches a human ear, it can cause a stimulus to reach the brain and hearing results. Thus, to the physicist, sound is the wave motion itself. Sound can also be defined as the sensation produced when sound waves reach the ear. Because both definitions are in common use, care must be taken to avoid confusion.

Light is the particular part of the gamut of electromagnetic waves to which the eye responds. Here, varying electric and magnetic fields travel through a vacuum or a transparent medium. The energy travels in even multiples of specific small amounts called "quanta." Related to light are infrared radiations, those frequencies which are just too low to cause the sensation of light in the normal human eye, and ultraviolet rays, similar radiations whose frequencies are just too high for the human eye. Optics, which is the study of light, often includes infrared and ultraviolet radiation.

Electricity and magnetism are so closely related that neither topic can be studied in any depth without involving the other. Electricity deals with

the forces which charged particles exert upon each other, the effects of such forces, and the phenomena caused by the motions of charged particles. Magnetism was first known in the form of the peculiar attraction which a mineral called lodestone exerted because of the particular electron orientations in the material. It was found that magnetic effects could be caused or altered by electric currents. Further, it is possible to use magnetic materials to advantage in producing electric currents, which are streams of charged particles. An important branch of electricity is electronics, and this area has attracted the attention of more physicists in this country than any other part of physics. This subject deals with the flow of electric currents in vacuum tubes, semiconductors, and associated circuits, and with the use of these circuits to form devices of many kinds.

Solid-state physics is a well established area cutting across some of the other areas mentioned above. It includes those phenomena which are exhibited by materials in the solid state, with emphasis on electronic properties and their relations to the composition of crystalline substances and to energy levels in these materials.

Modern physics is a title under which physicists group many topics of comparatively recent development involving or relating to atoms and other submolecular particles or radiation resulting from atomic processes. Subjects which are considered here include radioactivity (both natural and artificially induced), x-rays, atomic and molecular structure, the quantum theory, wave mechanics and matrix mechanics, and nuclear fission and fusion.

Experimental physics and theoretical physics are two general categories into which the entire field of physics can be divided. The experimentalist attempts to discover new phenomena through the manipulation of apparatus and to make measurements of old or newly discovered properties. The theoretical physicist attempts to correlate measurements and to simplify theories. He tries to predict new phenomena and to relate one effect to another. Mathematical physics uses mathematics to describe physical phenomena, and extends or adapts mathematics to make it applicable to specific theories.

General discussions of major areas of physics can be found in the cross-references listed at the end of this article.

Future Trends. It is of course impossible to predict the flow of any science as it varies from quiet pools to violent rapids. Present trends lead one to expect that certain areas will receive emphasis, but a new discovery or even gradual progress in an old area can open up vast new vistas. Computers and computer techniques are speeding calculations on nuclear forces, and some of the riddles of their existence may yield to intense pressures. Fundamental-particle physics is another region with many questions which have been posed but not answered. Which, if any, of the particles are truly fundamental? And what does "charged" imply when applied to particles?

What will even more powerful accelerators reveal? The emission of coherent radiation and its behavior combine to form an area in which effort is being accelerated. The Fermi surfaces of metals and superconductivity and other cryogenic phenomena are fruitful, lively and interesting areas. Undoubtedly more and more surprises are in store from expanding research programs.

Physics Organizations. Throughout the world there are many societies and other organizations that have the objectives of advancing physics and distributing knowledge of that science. In addition, other organizations deal with broader or related fields and include physics as part of their interests. In the United States of America, five societies—the Acoustical Society of America, the American Association of Physics Teachers, the American Physical Society, the Optical Society of America, and the Society of Rheology—are members of the American Institute of Physics, which was set up to assist them in their activities. Many other organizations are associated or affiliated with the Institute. It maintains an Information Center on International Physics Activities and publishes an "Information Booklet on Physics Organizations Abroad." Inquiries should be sent to the director of the Information Center, American Institute of Physics, 335 East 45th Street, New York, N.Y. 10017, U.S.A. In addition, scientific personnel attached to embassies in Washington, D.C. can provide information about activities in their respective countries.

ROBERT M. BESANÇON

Cross-references: ACOUSTICS; ASTROPHYSICS; ATOMIC PHYSICS; BIOPHYSICS; CHEMISTRY; ELECTRICITY; ELECTRONICS; GLOPHYSICS; GRAVITATION; HEAT; LIGHT; MAGNETISM; MATHEMATICAL PHYSICS; MEASUREMENTS, PRINCIPLE OF; MECHANICS; MOLECULES AND MOLECULAR STRUCTURE; PHYSICAL CHEMISTRY; RADIOACTIVITY; RELATIVITY; THEORETICAL PHYSICS; MATHEMATICAL PHYSICS.

PLANCK'S LAW. See **RADIATION, THERMAL.**

PLANETARY ATMOSPHERES

The atmospheres of the planets within our solar system are very different from each other; in fact, no two planets have identical atmospheres. Nevertheless, the planets can be divided into two major categories within which some similar characteristics are present. These two categories are the "terrestrial" planets, earth, Venus, Mars Mercury, which are fairly high-density solid planets with relatively thin atmospheres, and the "giant" planets, Jupiter, Saturn, Uranus, and Neptune, which have much lower densities and extensive atmospheres.

Within these categories, a distinction can be made between those characteristics such as temperature and concentration of impinging radiant energy, which are determined primarily by the planet's distance from the sun, and others, such as

escape velocity, which are determined by the mass of the planet itself. However, these distinctions are subject to variation, due to interaction of various environmental factors, and as a result, each planet must be considered individually for best understanding.

Our own planet, naturally, is the one about which we have the most extensive information and thus is the one which is best understood. Conclusions about the nature of other planetary atmospheres must often be made from very inadequate data and consequently may sometimes be speculative. Our major sources of information about the composition of other planetary atmospheres are the spectroscopic studies made by astronomers for many years. Tables 1 and 2 list important data for the various planets.

Terrestrial Planets. *Mercury* is the closest planet to the sun and, as expected, has the highest surface temperature. With its low mass, Mercury has a correspondingly low escape velocity and any atmosphere present in the past should have escaped to space. Since its period of rotation is equal to its period of revolution, Mercury always turns the same face toward the sun. The dark side, therefore, is very near absolute zero. Urey¹ has discussed the possibility that a very thin atmosphere of high-molecular-weight substances may exist and account for the transient haze sometimes observed on the planet.

The atmosphere of *Venus* is somewhat mysterious at present despite centuries of observation. The only major constituent which has been positively identified is CO₂, which is known from spectroscopic measurements to be present to the extent of 10⁵ cm-atm above the cloud layer. From cosmic abundance data nitrogen is also expected as a major constituent. Hydrogen and helium would have escaped to space over geologic ages. Trace amounts of CO and O₂ should be present from photodissociation of CO₂ in the upper atmosphere, and these gases have recently been observed spectroscopically in very low abundance. In the upper atmosphere, O₂ will be dissociated

into O atoms, and the CO and O atoms can recombine to regenerate CO₂ in a number of ways, for example:



Thus the dissociation products do not accumulate to any large extent. Water vapor has not been observed with certainty, although reports have been published of the possible presence of trace amounts. It seems unlikely that either atmospheric or surface water can be present in significant amounts for two reasons. The first is the high level of CO₂ in the Venus atmosphere. CO₂ is believed to have been removed from the earth's atmosphere by reaction with silicate rocks, a reaction which proceeds only in the presence of water. The second is the absence of high concentrations of oxygen in the Venus atmosphere. In the earth's atmosphere, oxygen is believed to have originated in part by photodissociation of H₂O and escape of H atoms to space.

The surface of the planet is completely obscured by highly reflecting clouds whose composition is unknown. Although it is often speculated that the clouds are water droplets or ice crystals, this seems unlikely at present. The possibility that the clouds could consist, at least in part, of carbon suboxide polymer has also been considered, since it is known that such polymers are formed by the action of ionizing radiation on both CO₂ and CO.

Until recently most astronomers believed, on the basis of infrared radiometer data and the distance of Venus from the Sun, that the temperature of Venus was about the same as that of the earth, or perhaps a little higher. In the past few years, radiometer data in the microwave region have been obtained which, if interpreted as thermal radiation, lead to the conclusion that the surface temperature are very high, of the order of 600 K or higher. (The temperature of the cloud layer, however, is much lower.) None of the theories advanced to account for such a high surface temperature are very convincing and it has

TABLE 1. PLANETARY DATA*

Planet or Satellite	Distance from Sun (earth = 1)	Mass (earth = 1)	Radius (earth = 1)	Temperature (black-body av) (°K)	Density (g/cc)	Period of Revolution	Period of Rotation	Escape Velocity (km/sec)
Mercury	0.3871	0.0543	0.39	443	5.05	88 days	88 days	4.3
Venus	0.7233	0.8136	0.973	262	4.88	224.7 days	224.7 days	10.4
Earth	1	1	1	250	5.52	365 days	24 hours	11.3
Mars	1.5237	0.1080	0.520	218	4.24	687 days	24 hours 37 minutes	5.1
Jupiter	5.2028	318.35	10.97	105	1.33	11.86 years	9 hours 55 minutes	61.0
Saturn	9.5388	95.3	9.03	78	0.714	29.5 years	10 hours 38 minutes	36.7
Uranus	19.19	14.54	3.72	55	1.56	84 years	10.8 hours	22.4
Neptune	30.07	17.26	3.50	43	2.22	165 years	16 hours	25.5
Pluto	39.52	0.033(?)	0.45(?)	42	2.0(?)	248 years	?	3(?)
Titan	9.5388	0.0236	0.384	83	2.30			2.8

* Data in the first five columns are taken from Urey's¹ article. The values in the last column are those given by Kuiper.²

TABLE 2. CONSTITUENTS OF PLANETARY ATMOSPHERES*†

Planet or Satellite	Constituent	Amount (NTP-earth)†† (cm-atm)	Remarks
Venus	CO ₂	100 000	Observed spectroscopically
	N ₂		Believed major constituent from cosmic abundances
	CO	< 100	Spectroscopic upper limit
	O ₂	< 100	" " "
	CH ₄	< 20	" " "
	N ₂ O	< 100	" " "
	NH ₃	< 4	" " "
Earth	N ₂	625 000	
	O ₂	168 000	
	Ar	7 400	
	CO ₂	200	
	CH ₄	1.2	
	O ₃	0.3	
	H ₂ O	Variable	
Mars	CO ₂	3 600	Observed spectroscopically
	N ₂	180 000	Calculated from total pressure, believed major constituent
	H ₂ O		Believed present in trace amounts
	CH ₄	< 10	Spectroscopic upper limit
	NH ₃	< 2	" " "
Jupiter	CH ₄	15 000	Observed spectroscopically
	NH ₃	700	
	H ₂	$2.7 \cdot 10^7$	Calculated by Urey ¹ from cosmic abundances; based on low density of planet
	He	$5.6 \cdot 10^6$	
	N ₂	$4 \cdot 10^3$	
	Ne	$1.7 \cdot 10^4$	
	CH ₄	35 000	Observed spectroscopically
Saturn	NH ₃	200	
	H ₂	$6.3 \cdot 10^7$	Calculated by Urey ¹ from cosmic abundances; based on low density of planet
	He	$1.3 \cdot 10^7$	
	N ₂	$9.5 \cdot 10^3$	
	Ne	$2.7 \cdot 10^4$	
	CH ₄	220 000	Observed spectroscopically
Uranus	H ₂	$4.2 \cdot 10^6$	H ₂ observed spectroscopically
	He	$8.6 \cdot 10^5$	Abundances calculated by Urey ¹
	N ₂	$4.2 \cdot 10^6$	
	CH ₄	370 000	Observed spectroscopically
Neptune	H ₂		Detected spectroscopically at higher intensity than Uranus
	N ₂		
	He		
	CH ₄	20 000	Observed spectroscopically
Titan	NH ₃	< 300	Spectroscopic upper limit

* Data taken from Urey¹ and Kuiper.²

† No constituents are listed for Mercury which has no atmosphere.

†† One cm-atm (NTP-earth) is an amount of gas equivalent to a column 1 cm in height at normal atmospheric pressure and temperature.

been suggested that the microwave radiation originates from a nonthermal source such as thunderstorms, chemical reactions in the atmosphere, or electrical charge fluctuations carried on atmospheric dust particles. Many detailed discussions have been published of the lower atmosphere of Venus, including temperature variations, pressure, height and thickness of the cloud layer, and relative abundance of atmospheric constituents. Which, if any, of these is correct can be decided only after more information becomes available.

Because of the obscuring cloud layer, no surface features of the planet can be observed and the period of rotation cannot be determined by the usual astronomical methods. However, from radar data and the absence of a strong magnetic field, as found by Mariner II, it appears very likely that Venus, like Mercury, has a period of rotation equal to its period of revolution, and thus may have one permanently dark side. If this is so, the dark side may be very cold, although atmospheric convection may provide a means of warming. Infrared radiometer data show that

the cloud layer temperature of the dark side is not substantially different from that of the bright side. Many questions obviously remain to be answered about this planet which in so many ways is the earth's "twin."

The atmosphere of our own planet, *earth*, has been the subject of many extensive studies for years and a great deal of information has been gathered. The major atmospheric constituents are nitrogen and oxygen in a 4:1 ratio. Other minor constituents include CO_2 , H_2O , and Ar. Hydrogen and helium can escape from the earth's atmosphere and over geologic ages would have disappeared. The variation of pressure with altitude follows the exponential equation

$$P = P_0 e^{-\mu g h / RT}$$

where μ is the average molecular weight and g is the acceleration due to gravity. The quantity $RT/\mu g$ is called the scale height and given the symbol H . This relationship is only approximately correct because neither μ or T is constant throughout the atmosphere. For very correct calculations, changes in g must also be considered. The earth's atmosphere shows an interesting variation of temperature with altitude which arises from several factors. The earth's atmosphere is transparent to most solar radiation, which penetrates to the earth's surface and heats it to a temperature of about 290 K. The earth radiates this energy back out to space, heating the atmosphere to an extent which decreases with increasing distance from the surface. The atmospheric temperature drops gradually to 217 K at 20 km. At about 20 km, the temperature of the atmosphere begins to show changes brought about by the basic photochemistry of the atmosphere. The temperature rises to a value of about 283 K at 50 km and then decreases gradually to 168 K at 80 km. Above this altitude the temperature rises steadily to about 1500 K at 400 km and above. This maximum changes with the sunspot activity and at sunspot minimum is 1000 K.⁴

The photochemical process contributing to these variations include the following. Ultraviolet light from the sun is absorbed by O_2 molecules in the upper atmosphere, resulting in their dissociation into oxygen atoms. The atoms may then react with other O_2 molecules to form ozone



The ozone absorbs light of wavelengths below 3000 Å very strongly and, in fact, protects the earth's inhabitants from this energetic radiation. The ozone is destroyed both by this absorption of light and by further reaction with atomic oxygen



with a steady-state equilibrium being set up between formation and destruction. This occurs in a relatively narrow band at ~30 km altitude, known as the ozone layer. The energy absorbed in this narrow layer is responsible for the temperature rise. At the highest altitudes very high

energy solar radiation is absorbed, producing ions of oxygen, nitrogen, and other gases. This region is known as the ionosphere. Detailed discussions of the physics and chemistry of the earth's atmosphere have been published by Kuiper,² and Bates³ as well as others. Very extensive studies of the earth's atmosphere have been carried out in recent years using rocket probes. Much of the information obtained in this way has been summarized by Johnson¹ in the "Satellite Environment Handbook."

The atmosphere of *Mars* is relatively thin, as might be expected from the low mass and consequent low escape velocity. Urey¹ notes that over geologic ages even oxygen and nitrogen atoms may have escaped from Mars in addition to hydrogen and helium if the upper atmosphere temperature is high enough. The only atmospheric constituent identified positively is CO_2 . The total atmospheric surface pressure has been estimated by several different workers to be of the order of 1/10 that of the Earth and it is believed that nitrogen is the most abundant constituent. The Martian surface seems to have polar frost caps which disappear in summer and there is some evidence of seasonal vegetation, both of which indicate the presence of water. However, this must be in small amounts since both H-atoms and O-atoms will escape from the Martian atmosphere. Clouds are observed in the atmosphere which many conclude are water clouds. In addition a blue haze is sometimes present, whose origin is the subject of many discussions in the literature. It may be due to dust particles or to fluorescence radiation of atmospheric ions or to some other cause at present unknown. It seems improbable that oxides of nitrogen can be present in the Martian atmosphere, as recently reported, for the following reasons. Kinetic studies on the fixation of nitrogen by ionizing radiation have shown that such compounds do not form in appreciable amounts at these pressures. In addition, any small amounts formed would be destroyed by solar ultraviolet radiation.

Giant Planets. Because of their high masses and consequent high escape velocities the atmospheres of the giant planets differ substantially from those of the terrestrial planets. Hydrogen and helium cannot escape from the giant planets and, on the basis of cosmic abundance, can be expected as the major atmospheric constituents. Supporting this expectation of a reducing atmosphere, is the observed presence of a large amounts of methane in the atmospheres of all these planets and the mean molecular weight of 2.35 calculated for Jupiter from the attenuation of the atmosphere.

The atmosphere of *Jupiter* contains appreciable amounts of ammonia (700 cm-atm) in addition to an observed 15 000 cm-atm of CH_4 . Urey¹ has calculated that 2.7×10^7 cm-atm of hydrogen and 5.6×10^6 cm-atm of helium should also be present. Many colors are observed in the Jupiter atmosphere, blues, greens, yellows, and reds. Often distinct bands appear. Rice⁵ has suggested that these colors may be caused by free radicals

formed photochemically and stable at the low ambient temperatures. The most notable feature of the Jupiter atmosphere is the Great Red Spot. This unusual formation floats (in the atmosphere?) and is not rigidly attached. No satisfactory explanation has been proposed for its nature or origin. The turbulence of the cloud bands indicates that very stormy conditions prevail in the atmosphere. Recent microwave radiometer data indicate that Jupiter is surrounded by an intense radiation belt similar to the van Allen belt.

The atmosphere of *Saturn* appears to be similar in nature to that of Jupiter, with 35 000 cm-atm of methane and 200 cm-atm of ammonia being observed spectroscopically. From cosmic abundances 6.3×10^7 and 1.3×10^7 cm-atm of hydrogen and helium have been calculated by Urey.¹ Neon and nitrogen should also be expected. Colors and bands are observed in the atmosphere but they are much less pronounced than those of Jupiter. The rings of Saturn, one of the most striking features of the solar system, are believed by Kuiper² to be ice particles.

Uranus and *Neptune*, because of their high densities, appear to have different compositions, specifically, to contain smaller proportions of hydrogen and helium than do Jupiter and Saturn.¹ Ammonia has not been detected in either planet, but Kuiper² reports 220 000 and 370 000 cm-atm of methane on Uranus and Neptune respectively. The (r^1 , r^{11}) (3,0) pressure induced rotation-vibration band of H₂ has also been identified by Herzberg (see references 1 and 2). Several other spectroscopic bands have been observed which have not been positively identified to date. One band has been attributed to HNO, but it seems unlikely that this compound can be present to a significant extent.

Almost nothing is known about *Pluto* which because of its great distance from the sun, is very cold. No atmospheric constituents have been detected. This may be due to the fact that hydrogen and helium can escape and other substances except neon have no appreciable vapor pressure at this low temperature of 42°K.

Jupiter, Saturn, and Neptune all have fairly large satellites but of these only Titan, a satellite of Saturn, has a detectable atmosphere. Methane (20 000 cm-atm) is the only constituent observed.

P. HARTECK
B. A. THOMPSON

References

1. Urey, H. C., "The Atmospheres of Planets," "Handbuch der Physik," Vol. 52, pp. 363-418, Berlin, Springer Verlag, 1959.
2. Kuiper, G. P., Ed., "The Atmospheres of the Earth and Planets," Chicago, University of Chicago Press, 1952.
3. Bates, D. R., "The Earth and Its Atmosphere," New York, Basic Books, 1957.
4. Johnson, F. S., Ed., "Satellite Environment Handbook," Stanford, California, Stanford University Press, 1961.
5. Rice, F. O., *Sci. Am.*, **194**, No. 6, 119 (1956).

Cross-references; CLOUD PHYSICS, IONOSPHERE, RADIATION BELTS.

PLASMAS

When a gas is raised to sufficiently high temperatures, the atoms and molecules of the gas become ionized since electrons are stripped off by the more violent collisions ensuing from the increased thermal agitation of the particles. In the resultant highly ionized state, the dynamical behavior of the gas can be dominated by the electromagnetic forces acting on the now unbound electrons and ions, and the properties become sufficiently different from those of the normal un-ionized gas to warrant a new name, *plasma*, to describe the gas in this highly ionized state. This name arose from the expression "plasma oscillation" coined by Langmuir to describe the very high frequency (1000 Mc/sec) longitudinal oscillations sustained by plasma in regimes where normal sound waves are characteristically strongly damped.

The modern studies of molecular, atomic, and nuclear physics had their origins in the study of the conduction of electricity through gases, and indeed, until quite recently, the (weakly) ionized gas has been used primarily to study atomic structure and the complex of collision processes—ionization, excitation, inelastic and resonant scattering, etc.—which occur when an electron strikes an atom, rather than the plasma properties (see COLLISIONS OF PARTICLES AND IONIZATION). The highly ionized plasma, with its dominant collective behavior, can be produced and observed in the laboratory only with considerable difficulty, and it is not surprising that the study of ionized gases was largely the study of atomic collision processes.

Largely due to the astrophysicist, the interest in the plasma *per se* has been revived in recent years. It was realized that the plasma state was essentially the normal state of matter in this universe. Hydrogen, which is overwhelmingly the most abundant element in the stars and space, is mostly ionized and hence totally stripped. Helium, the next most abundant element is stripped of its two electrons inside the sun and in the solar corona. Nearer home, the properties of the solar wind, Van Allen belts, and the ionosphere, demand a plasma description.

Most recently, the technological possibilities of plasma have attracted attention, most significantly in the research directed toward the controlled release of energy from the thermonuclear fusion of light elements, and it was with this stimulus that a considerable advance in the description and understanding of plasma phenomena took place. The possibilities of ionic propulsion for interplanetary flight and schemes for direct thermionic conversion have also attracted considerable attention.

We thus see that the study of the ideal plasma, a gas composed entirely of electrons and bare ions in which the inelastic atomic processes of

normal gaseous electronics are unimportant, is a subject worthy of investigation.

The problems encountered in analyzing a fully ionized plasma are of several types. Although the basic physical processes are simpler than in an ordinary gas, the motions are vastly more complicated because of the strong coupling to the electromagnetic field. From a macroscopic point of view, the plasma interacting with a strong magnetic field can often be considered, in a certain approximation, as a highly electrically conducting fluid, and thus described by a combination of hydrodynamic equations for the fluid and Maxwell's equation for the electromagnetic field. As expected, the resultant motions yield a vastly richer variety of flows than encountered in ordinary hydrodynamics and these analyses have developed into a separate discipline called *magnetohydrodynamics*, *hydromagnetics*, or **MAGNETO-FLUID-MECHANICS**.

Even in the absence of a magnetic field, the electrical properties of plasma permit complicated macroscopic motions, which involve electrostatic restoring forces and which have no parallel in ordinary gases.

From a microscopic point of view, the plasma consists of an assembly of charged particles interacting under simple inverse-square forces (and any external electromagnetic fields) and appears as an ideal subject for classical kinetic theory. However, because of the long-range character of the coulomb force, the simple picture of a diffuse gas in which the interaction can be represented as a scarce, catastrophic, binary encounter, is not valid. The microscopic dynamics of a plasma must be properly treated as a problem in many-body physics, and it is only quite recently that it has been learned how to do this in a systematic way.

The statistical many-body description of plasma may be characterized by the orders in an expansion in the parameter $\epsilon = 1/n\lambda_D^3$, where n is the average number density of particles, and λ_D , the Debye shielding distance, is given by

$$\lambda_D = [4\pi n e^2 (1 + z)/\Theta]^{-1/2}$$

Here, e is the absolute value of the electronic charge, z is the atomic number, and $\Theta = kT$ where T is the absolute temperature and k is Boltzmann's constant.*

The parameter ϵ , the inverse of the number of particles in a Debye sphere, is small and ranges, typically, from 10^{-2} to 10^{-9} ; indeed, the smallness of this parameter typifies the plasma state. The Debye length characterizes the distance over which there can be considerable departure from average charge neutrality. It measures not only the thickness of the charged sheath which forms when plasma is in contact with a solid surface, but also the effective range of the potential due to a singled-out test charge at rest

$$\phi(r) = \frac{q}{r} e^{-r/\lambda_D}$$

* Unrationalized gaussian (cgs) units are used throughout this article.

where r is the distance from the charge q . The exponential decay factor represents the average collective many-body shielding effect.

There is an important interpretation of the smallness of the plasma number ϵ . If one computes the ratio of the potential energy of interaction of one particle with its neighbor and compares it with the mean kinetic energy per particle, one finds

$$PE/KE \sim \frac{e^2 n^{1/3}}{\Theta} \sim (n\lambda_D^3)^{-2/3} \ll 1$$

Thus the effect of any one particle on a given particle is small compared to the average collective many-body effect of all its neighbors within a Debye sphere.

The average collective behavior of plasma is described (to zero order in ϵ) by the so-called Vlasov or self-consistent approximation for the density in the six-dimensional r, v phase space. The density (or one-particle distribution function) $f_s(r, v, t)$ satisfies the equation for each species s

$$\frac{df_s}{dt} + \mathbf{v} \cdot \frac{\partial f_s}{\partial \mathbf{x}} + \frac{e_s}{m_s} \left[\mathbf{E}(\mathbf{r}, t) + \frac{\mathbf{v}}{c} \times \mathbf{B}(\mathbf{r}, t) \right] \cdot \frac{\partial f_s}{\partial \mathbf{v}} + \frac{\mathbf{F}_{\text{ext}}}{m_s} \cdot \frac{\partial f_s}{\partial \mathbf{v}} = 0 \quad (1)$$

where $e_s = ze$ - ionic charge of species s , m_s is the ionic mass of species s and c is the velocity of light. \mathbf{E} and \mathbf{B} satisfy Maxwell's equations with sources

$$\sigma(\mathbf{r}, t) = \sum_s e_s \int d^3v f_s$$

and

$$\mathbf{j}(\mathbf{r}, t) = \sum_s e_s \int d^3v \mathbf{v} f_s$$

Here σ and \mathbf{j} are the average charge and current density respectively. \mathbf{F}_{ext} represents any external force field which may be present.

For undriven systems which are spatially homogeneous on the average $\mathbf{E} = \mathbf{B} = 0$, hence $\partial f / \partial t = 0$, and one must proceed to next order in the plasma number, where interaction (collisions) between pairs of *shielded* particles is considered, in order to discuss the time evolution of f .

It is observed that, in this lowest approximation, any f_0 which is a function of v alone satisfies this set of equations for the spatially homogeneous state. If one considers small departures from the spatially homogeneous situation with $f_0 = f_0(v)$ one finds the plasma capable of sustaining high-frequency, long-wavelength, plane-wave excitations of two types: longitudinal, ($\mathbf{k} \parallel \mathbf{E}$), with

$$\omega^2 \approx \omega_p^2 + 3 \frac{\Theta}{m} k^2$$

and transverse, ($\mathbf{k} \perp \mathbf{E}$), with

$$\omega^2 \approx \omega_p^2 + c^2 k^2$$

Here $\omega_p = (4\pi n e^2 / m)^{1/2}$ is the *electron plasma frequency* and m is the mass of the electron. The

much more massive ions play no role in this high-frequency situation. The electrostatic (Langmuir) oscillation actually are found to be damped, even in this collisionless approximation, by the interaction of the wave with particles traveling near the phase velocity of the wave, "Landau Damping." If the ion temperature is small compared to the electron temperature, there also exist low-frequency ion excitations in which the electrons play a quasi-static role.

The presence of a uniform magnetic field greatly increases the variety of excitations possible, but these will not be discussed here.

If the distribution f_0 is sufficiently anisotropic in velocity space, it is possible under certain circumstances, to find growing excitations. The ultimate fate of these oscillations cannot be described by the linearized theory, and a nonlinear treatment must be invoked. If the growth rate of these excitations is small, however, a "quasi-linear" or adiabatic treatment is justified and such investigations are currently very much in vogue with plasma physicists. In the presence of an external magnetic field, even small average density and temperature gradients can lead to unstable situations.

In the next order in ϵ , the effects of particle discreteness (collisions) come into play, and here one meets the kinetic theory of plasma. The relaxation to thermal equilibrium is described by a modified *Fokker-Planck* collision term which is added to the lowest order description¹ and which is characterized as a superposition of weak, statistically independent, but dynamically shielded, binary encounters. A novel addition to the usual type dissipative collisional mechanism is the emission of the forementioned longitudinal electrostatic oscillations by superthermal particles.

An interesting and important property of plasmas is their ability to radiate. At the Vlasov level of description, a spatially bounded plasma in a state of longitudinal excitation can emit transverse waves (very much like an antenna!).

To first order in ϵ , and in the presence of a magnetic field, the acceleration of charged particles in their spiraling motion about the magnetic field lines gives rise to radiation at the gyro-frequency, $\omega_c = eB/mc$, and its harmonics. These lines are not sharp, however, due to the Doppler effect of motion along the field lines, and the change in frequency due to the relativistic mass correction.

To second order in ϵ , the radiation from the particle acceleration involved in binary collisions, *bremsstrahlung*, is described.

These radiative loss mechanisms represent serious competition for the energy gain in the efforts for the controlled release of energy through the fusion of light elements.

CARL OBIRMAN

References

- Spitzer, L., Jr., "Physics of Fully Ionized Gases," New York, Interscience Publishers, Inc., 1956.

Stix, T. H., "Theory of Plasma Waves," New York, McGraw-Hill Book Co., 1962.

Thompson, W. B., "Introduction to Plasma Physics," London, Pergamon Press, 1962.

Cross-references: ASTROPHYSICS, COLLISIONS OF PARTICLES, IONIZATION, IONOSPHERE, KINETIC THEORY, MAGNETO-FLUID-MECHANICS, RADIATION BELTS.

POLAR MOLECULES

The term "polar" is applied to molecules in which there exists a permanent spatial separation of the centroids of positive and negative charge, or dipole moment. Such a moment was first postulated by P. Debye for molecules having structural asymmetry in order to explain certain of the observed electrical properties. He chose for his model an electrical dipole contained in a spherical molecule, free to rotate into alignment with an applied electric field, but subject to disorientation by collisions with other molecules due to thermal motion. At ordinary temperatures and field strengths, the electrical energy involved in the orientation is much smaller than the thermal energy kT so that only a small fraction of the dipoles are aligned with the field (k is Boltzmann's constant and T is the temperature in degrees Kelvin). The net dipole moment per mole (*molar polarizability*) is then calculated statistically to be

$$P = \frac{4\pi}{3} N \left(\alpha + \frac{\mu^2}{3kT} \right) \quad (1)$$

where μ is the permanent dipole moment per molecule, i.e., the charge multiplied by the distance of separation, and N is Avogadro's number. The polarizability α represents the induced moment per molecule resulting from the temporary distortion of the electron orbits by the applied field. For nonpolar molecules, it is the only contribution; it corresponds to the optical polarizability as measured by the refractive index.

The molar polarizability may be related to the dielectric constant ϵ by the approximate equation of Clausius and Mosotti

$$\frac{\epsilon - 1}{\epsilon + 2} \frac{M}{d} = P \quad (2)$$

where M is the molecular weight and d is the density. The permanent dipole moment of a molecule may therefore be determined by measuring the temperature coefficient of the dielectric constant as seen by combining Eqs. (1) and (2). Alternatively, a companion measurement of the optical refractive index may be made to determine α , and the dipole moment is then obtained as the difference between the total polarization and the optical contribution. The measurements are made in dilute vapor or solution phase, so that the individual dipoles are sufficiently far apart that they do not influence one another (see REFRACTION).

Equation (1) is valid in the low-frequency region where the dipoles are able to rotate in phase with the applied electric field. This rotation is subject to various restraints; in the simple case of a spherical molecule of radius a rotating in a fluid of viscosity η , this leads to a relaxation time

$$\tau = \frac{4\pi a^3 \eta}{kT} \quad (3)$$

corresponding to a frequency range of *anomalous dispersion* in which the molecules become unable to follow the oscillations of the applied field. This gives rise to an out-of-phase component of the dielectric constant representing a conductivity or *dielectric loss*, ϵ'' , i.e., a dissipation of energy in the form of heat. Mathematically this is expressed as a complex dielectric constant.

$$\epsilon = \epsilon' - j\epsilon'' \quad (4)$$

$$\epsilon' - \epsilon_\infty = \frac{\epsilon_0' - \epsilon_\infty}{1 + \omega^2 \tau^2} \quad (5)$$

$$\epsilon'' = \frac{(\epsilon_0' - \epsilon_\infty) \omega \tau}{1 + \omega^2 \tau^2} \quad (6)$$

ω is $2\pi \times$ the frequency, and the subscripts 0 and ∞ refer to dielectric constant measured at very low and very high frequency, respectively. Cole and Cole showed for polar molecules having a unique relaxation time that a plot ϵ'' vs ϵ' is a semicircle centered on the ϵ' -axis. For more complicated molecules or high polymers, the center becomes depressed below the ϵ' -axis and the curve may be further distorted. This behavior is generally characterized by a distribution of relaxation times as a result of the orientation of molecular segments of various shapes and sizes. Information regarding the freedom of orientation within the molecules may thus be gained.

These simple relationships between dipole moment and dielectric constant fail for concentrated solutions or pure polar liquids because they do not take into account the interaction of the dipoles, both permanent and induced, with one another. The calculations have been extended by Onsager, Kirkwood, and others, to include these effects. It becomes necessary to include in the theory a correlation factor which is a measure of the extent of nonrandom orientation of the dipoles, i.e., the tendency of the dipoles to aggregate parallel or antiparallel to one another. Experimental determination of this quantity yields further insight into the structure of polar liquids.

In the solid state, most crystalline polar compounds exhibit low dielectric constants because the rotational freedom necessary for dipole orientation has been frozen out. A few compounds, however, do show a persistence of high dielectric constant to temperatures below the melting point, indicating rotational freedom in the solid. At some lower temperature the rotation ceases

and the dielectric constant drops. This change in dielectric constant has frequently been used as a method of detecting second-order transitions in polar compounds.

Some success in the correlation of dipole moment with molecular structure has been achieved. A series of moments assigned to individual bonds was developed empirically by Smyth, Pauling, and others from measurements on simple molecules. They are intended for approximate calculation of dipole moments of complex molecules by vectorial addition along the bond directions as determined by other means. It is assumed that there is no interaction between bonds; this generally results in the calculated values being higher than the measured moment because of inductive effects, i.e., electrons being shared in a bond between two atoms are not wholly available to a neighboring bond, so that neither bond attains its full moment. An approximate correction for this effect has been made by Eyring and co-workers. Despite these inadequacies, the calculated moments are often helpful in determining molecular structures and have often applied in the case of various substituted benzene-ring compounds. Of special importance is the ability to decide between a polar or nonpolar, i.e., symmetrical, structure.

More detailed quantum mechanical calculations of dipole moments have been less successful. Rather the experimentally determined moments have been used to assign varying degrees of ionicity and covalency to the bonds in establishing the electronic hybridization structure. Quantum mechanics has also shown how it is possible to obtain extremely accurate measurements of the dipole moments from spectroscopic Stark splittings.

D. EDELSON

Cross-references: DIPOLE MOMENTS, DIELECTRIC THEORY, MOLECULES AND MOLECULAR STRUCTURE, POLYMER PHYSICS, QUANTUM THEORY, REFRACTION, ZEEMAN AND STARK EFFECTS.

POLARIZED LIGHT

Polarized light is an especially simple form of light and can be defined easily in terms of either of the two prevalent theories of light: the wave theory and the photon theory (see LIGHT). According to the former, light consists of trains of electromagnetic waves whose wavelengths lie in the range from about 4×10^{-7} to 7×10^{-7} meters. A noteworthy feature of the waves is that two kinds of displacements are involved: electric and magnetic. Another significant feature is that, when the waves are traveling in empty space (or in glass, water, or other isotropic medium) the electric and magnetic displacements are perpendicular to the direction of propagation of energy; i.e., they are transverse, not longitudinal. Since the electric and magnetic displacements are always perpendicular to one another and have equivalent magnitudes, it is sufficient to specify just one of these quantities; most

authors choose to deal with the electric displacement.

Consider, now, a slender beam of light that is traveling east, i.e., from left to right in Fig. 1. If it happens that the direction of electric displacement is everywhere up or down, but not north or south, the beam is said to be linearly polarized in the vertical plane. If, alternatively, the displacements were north and south, the beam would be called *horizontally* linearly polarized. The displacement might, of course, lie in some tilted plane specified by an angle α ; in such case, any given displacement may be regarded as the resultant of a vertical displacement and a horizontal displacement. The *sectional pattern* of any such beam is conventionally indicated by a straight line segment at the appropriate azimuth α , as indicated in Fig. 2(c). Linearly polarized beams that have azimuths differing by 90° are said to have *orthogonal* polarization forms.

Circular and elliptical polarization forms exist also. Circularly polarized light may be regarded as the result of combining two linearly

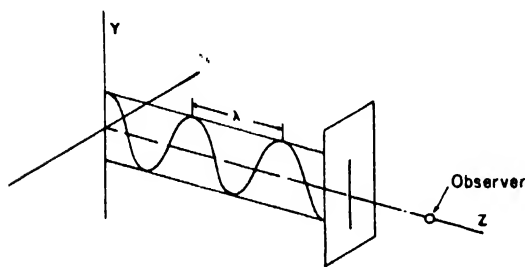


FIG. 1. Monochromatic beam of wavelength λ traveling horizontally to the right. Since, in this example, the electric displacement is vertical, the sectional pattern (indicated at right) consists of a vertical line and the beam is said to be vertically linearly polarized. The observer is situated far to the right, facing the light source.

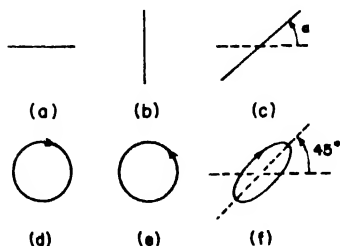


FIG. 2. Variety of sectional patterns of polarized light: (a) horizontally polarized; (b) vertically polarized; (c) linearly polarized at azimuth α ; (d) right circularly polarized; (e) left circularly polarized; (f) right elliptically polarized at 45° azimuth and with ellipticity, or ratio of semi-axes, of approximately 3.

polarized beams that have the same wavelength and same intensity, are polarized in orthogonal directions (e.g., horizontal and vertical) and differ in phase by $1/4$ cycle, or 90° . If the horizontally polarized component *lags* in phase by 90° , the sectional pattern of the combined beam is drawn as a circle executed clockwise, as judged by an observer facing towards the light source. If the horizontally polarized component *leads* by 90° , the circle has a counterclockwise sense and the light is called left circularly polarized. In the general case, the components may differ in magnitude and the phase difference may have any value; the general sectional pattern is an ellipse.

A different description of polarized light is required when the light is regarded as a stream of photons. This is the case when the photons are so infrequent and so energetic that they can be detected individually. In such case, it is the spin, i.e., the angular momentum, of the photon that constitutes the polarization. A right circularly polarized photon has a spin of $+1$ unit, and a left circularly polarized photon has a spin of -1 unit. Theory and experiment are in agreement that the magnitude of the unit in question is $h/2\pi$, where h is Planck's constant (see QUANTUM THEORY). Linearly and elliptically polarized photons may be regarded as combinations, in suitable proportions, of positive and negative spins.

Unpolarized light is more complex than polarized light, hence it is harder to describe. It consists of light in which the azimuth, ellipticity, and handedness of polarization vary rapidly and at random, so that no one type of polarization predominates. No simple diagram can depict the chaos and impartiality of the sectional pattern. If, in a given beam, one particular sectional pattern slightly outweighs all other patterns, the beam is said to be partially polarized; the degree of polarization may lie anywhere in the range from 0 to 100 per cent.

Polarization is not confined to visible light, but applies also to longer-wavelength radiations including the infrared and radio ranges and to shorter-wavelength radiations including ultraviolet light and x-rays. Nevertheless, the visual range deserves special attention in view of the variety of phenomena observed and the nicety of observation and wealth of applications.

Polarization was discovered by the Dutch scientist Christian Huygens in 1690, but it was not well understood until the transverse nature of the vibration, suggested by the English physicist Robert Hooke in 1757, was confirmed by Thomas Young in 1817. Further clarification came in 1873 when Maxwell showed that light waves belong to the family of electromagnetic waves.

Today many simple methods of producing polarized light are known. Usually, an investigator starts with a beam of unpolarized light and then polarizes it by inserting a suitable optical device—a polarizer. However, the invention of the LASER makes it feasible to generate light that

is polarized from the outset; no polarizer is needed. Various natural sources of polarized light exist, e.g. rays of light from a portion of the blue sky that is viewed in direction at 90° to the direction of the sun, also light from certain distant galaxies, such as the Crab nebula.

Conceptually, the simplest polarizer is the *micro-wire grid*, which consists of an array of parallel metallic wires each of which is less than a wavelength in diameter and is separated from its neighbors by comparably small distances. When a beam of unpolarized light strikes the grid, the component of the electric vibration that is perpendicular to the wires passes through readily, while the component that is parallel to the wires induces electric currents along them and is reflected or absorbed. The transmission axis of the device, defined in terms of the electric vibration of the transmitted component, is perpendicular to the wires. A far more economical type of polarizer is one that employs long, thin absorbing molecules, rather than wires. The most popular of the commercially produced polarizers, called H-sheet, contains large numbers of long, thin polymeric molecules consisting mainly of iodine atoms. These molecules are embedded in a plastic film that has previously been stretched unidirectionally so as to have a pronounced "grain"; the long slender molecules of iodine conform to this grain. Polarizers containing small absorbing units (whether wires or molecules) that show markedly different extents of absorption for different directions of electric vibration in the incident beam are called *dichroic* polarizers.

The first highly efficient polarizer was of birefringent type: it was made of the crystal *calcite* ($\text{CaO} \cdot \text{CO}_2$) which has two refractive indices and thus divides any incident beam into two beams. Each of these is linearly polarized, and the sectional patterns are orthogonal. Usually the crystal is artificially shaped so that one of the polarized beams is transmitted straight ahead and the other is deviated and disposed of by total internal reflection. The calcite prism designed by the Scottish physicist Nicol in 1828 was a basic piece of equipment in optics laboratories for over a hundred years. Types devised subsequently by Wollaston, Ahrens, and Foucault have proven to be superior in several respects.

Polarizers of reflection type are also well known. A typical reflection polarizer consists of a plate of glass that is mounted obliquely in the given beam of unpolarized light. The component that is transmitted is found to be partially polarized, and the reflected component is even more highly polarized—with the orthogonal sectional pattern. If the obliquity of the incident beam (measured from the normal to the plate) corresponds to *Brewster's angle*, defined as the angle that has a tangent of n , where n is the refractive index of the plate, the reflected beam is 100 per cent polarized. For ordinary glass, n has the value 1.5 and Brewster's angle is about 56° . Polarization by reflection is of common occurrence; yet few persons are aware of it; they are

unaware, for example, that light reflected obliquely from the surface of a pond, a wet road, or a glossy sheet of paper is partially linearly polarized.

Asymmetric scattering is another process that polarizes light. The polarization observed in light from the blue sky is a consequence of scattering of the sun's rays, especially the short-wavelength or blue component thereof, by the molecules of the air.

Perhaps the simplest application of polarizers is in controlling the intensity of a light beam. For this purpose two polarizers are used, in series. If they are oriented so that their transmission axes are parallel to one another, the over-all transmittance is large. If they are oriented so that the two axes are crossed, i.e., at an angle $\theta = 90^\circ$ to one another, the transmittance is zero; the beam is said to be extinguished. For intermediate angles-of-crossing the transmission is easily calculated from Malus' law, which affirms that the transmittance is proportional to $(\cos \theta)^2$. If the incident beam is already polarized, a single polarizer suffices to reduce the intensity to any desired extent. Polarizing sunglasses are effective in this manner thanks to the fact that light reflected obliquely from roads and most other nearly horizontal surfaces is partially *horizontally* polarized; since the polarizer lenses of the sunglasses are oriented with their transmission axes *vertical*, much of this reflected "glare" light is blocked. A polarizer that is employed to block an already polarized beam is called an analyzer.

Much of the interest in polarized light stems from the surprising convertibility of polarization form. By interposing an appropriate retardation plate, or *retarder*, in a given polarized beam, an experimenter can alter the polarization form at will, and with almost 100 per cent efficiency. A typical retarder consists of a thin flat crystal that exhibits birefringence, i.e., has two different refractive indices. Mica, being birefringent and being easily cleaved into thin plates, is often used. When a beam of polarized light enters a plate of mica, the beam is divided into two components and the phases of these are affected ("retarded") to different extents; thus when the components emerge from the plate and unite to form a single beam again, this latter is found to have a drastically altered sectional pattern. Especially versatile and accurate control of polarization form can be achieved with retarders of calcite or quartz. The effect of any given retarder on any given beam can be predicted accurately by various conventional means and, more recently, by a matrix algebra perfected by the American scientist Hans Mueller; the procedure is to multiply the four-element vector representing the beam by the sixteen-element matrix representing the retarder. Since the vectors and matrices are tabulated in various books, the procedure entails little effort; indeed it is readily extended to cases in which there are several retarders arranged in series. In cases where high accuracy is not needed, predictions can be made

especially rapidly with the aid of a kind of map, or spherical slide-rule, called the Poincaré sphere after its inventor Henri Poincaré.

When an object of glass or transparent plastic is subjected to a unidirectional stretching or compressing force, it becomes birefringent and thus acts like a retarder. Accordingly an engineer who wishes to evaluate the unidirectional strain within such an object can do so by directing a beam of polarized light through it and, with the aid of a calibrated retarder and an analyzer, measuring the change in the sectional pattern of the beam. Using conversion factors published in books on photoelastic analysis, he can interpret the change in terms of the direction and magnitude of the strain (see PHOTOELASTICITY).

Many microscopic biological objects, such as components of living cells, appear transparent and virtually invisible under a microscope. Yet such components often contain groups of aligned birefringent molecules and thus are capable of acting like miniature retarders. If the biologist illuminates a living cell with polarized light and examines it under a microscope that is equipped with an analyzer, he finds the birefringent components to be highly visible. Thus the use of polarizers renders visible a microscopic world that is normally invisible. Similarly, mineralogists find that the polarizing microscope greatly increases the visibility of small birefringent crystals.

There are many other applications of polarizers. Photographers use them to increase the contrast between white clouds and the (polarized) blue sky. Chemists use them to measure the extents to which various liquid solutions rotate the sectional pattern of a linearly polarized beam; the extent of rotation is a measure of the concentration of the solution. Electronics engineers use circular polarizers to trap and thus eliminate reflected glare from radar screens. Illumination engineers have devised systems of polarizing filters for automobile headlights and windshields that eliminate glare in nighttime driving. Biologists have found that the direction of growth of certain algae can be controlled by illuminating the algae with polarized light of controlled sectional pattern. Bees and ants can detect linear polarization directly by eye, and they employ the polarization of blue sky light as a navigational aid. Even man can learn to detect the polarization of white light with the naked eye, and he can also distinguish right from left circularly polarized light.

Physicists use polarizers to study the emission of polarized light by atoms situated in regions of strong electric or magnetic field, to determine the strength of the sun's magnetic field by measuring the polarization of certain solar spectral lines, and to determine the pattern of magnetic fields within the Crab Nebula. They use polarizers to analyze the behavior of the remarkable light source *the laser* and to verify theoretical predictions as to the polarization inherent in the synchrotron radiation emitted by certain high-energy accelerators.

Because of light's puzzling dual character (waves and photons) and its central position in the growing field of physics, and because *polarized* light is a most elemental form of light, physicists are confident that polarized light will continue to be an outstandingly challenging enigma as well as a most versatile tool for many generations to come.

WILLIAM A. SHURCLIFF

References

- Ditchburn, R. W., "Light," Second edition, New York, Interscience Publishers, 1963, 833 pp.
 Jenkins, F. A., and White H. E., "Fundamentals of Optics," Third edition, New York, McGraw-Hill Book Co., 1957, 639 pp.
 Land, E. H., "Some Aspects of the Development of Sheet Polarizers," *J. Opt. Soc. Am.*, **41**, 957 (1951).
 Shurcliff, W. A., "Polarized Light: Production and Use," Cambridge, Mass., Harvard University Press, 1962, 207 pp.
 Shurcliff, W. A., and Ballard, S. S., "Polarized Light," Princeton, N. J., D. Van Nostrand Co., 1964

Cross-references: ELECTROMAGNETIC THEORY, LASER, LIGHT, PHOTOELASTICITY, PHOTON, QUANTUM THEORY.

POLARON

An electron in the conduction band of an insulator (or a hole in the valence band) polarizes the medium in its neighborhood—this effect is particularly important in ionic crystals, on account of their high polarizability. The name "polaron" is given to the electron together with its associated cloud of lattice polarization. The subject has been of interest since Landau¹ suggested that an electron could become self-trapped by the lattice polarization it induces; this would mean that its effective mass would be very large compared with its "bare" mass (as given by band theory).

Frohlich, Pelzer and Zienau² have proposed a simple model Hamiltonian for the polaron. They assume that the ionic medium can be represented by a set of harmonic oscillators. Each oscillator is characterized by a wave vector \mathbf{q} , but the frequency ω is simply related to the Reststrahl frequency ω_r , and does not depend on \mathbf{q} . If \mathbf{p} is the electron momentum and \mathbf{r} its position, m the band mass, $\sigma_{\mathbf{q}}$ and $\sigma_{\mathbf{q}}$ canonical creation and annihilation operators for oscillator quanta ("optical phonons"), and Ω is the normalization volume, then the Hamiltonian is

$$H = -\frac{1}{2}\nabla^2 + \sum_{\mathbf{q}} \sigma_{\mathbf{q}}^{\dagger} \sigma_{\mathbf{q}} + i \left(\frac{2\sqrt{2}\pi\alpha}{\Omega} \right)^{\dagger} \sum_{\mathbf{q}} (\sigma_{\mathbf{q}}^{\dagger} e^{-i\mathbf{q}\cdot\mathbf{r}} - \sigma_{\mathbf{q}} e^{i\mathbf{q}\cdot\mathbf{r}}) \quad (1)$$

in units such that the electron band mass $m = 1$, $\hbar = 1$, and $\omega = 1$. The parameter α is defined as

$$\alpha = \frac{e^2}{\hbar} \left(\frac{1}{\epsilon_{\infty}} - \frac{1}{\epsilon_0} \right) \sqrt{\frac{m}{2\hbar\omega}} \quad (2)$$

with ϵ_∞ the frequency-dependent dielectric constant.

The theory, as so far formulated, is one of a particle interacting in a very simple way with a (nonrelativistic) field—the only parameter is the dimensionless coupling constant α . The problem is therefore of fundamental field-theoretical interest, as well as being interesting for its applications.

For very small α ($\alpha \leq 1$), a perturbation-theoretical and a Tamm-Dancoff type of variational calculation agree that Landau self-trapping does *not* occur.² For rather larger α ("intermediate coupling", $\alpha \leq 5$), the canonical transformation of Lee, Low and Pines,³

$$\left. \begin{aligned} \alpha_{\mathbf{q}}' &= (\alpha_{\mathbf{q}} + i c_{\mathbf{q}}) e^{-i \mathbf{q} \cdot \mathbf{r}} \\ \alpha_{\mathbf{q}}^{*'} &= (\alpha_{\mathbf{q}}' - i c_{\mathbf{q}}) e^{i \mathbf{q} \cdot \mathbf{r}} \\ \mathbf{p}' &= \mathbf{p} + \sum_{\mathbf{q}} \hbar \mathbf{q} \alpha_{\mathbf{q}}^{*'} \alpha_{\mathbf{q}} \\ \mathbf{r}' &= \mathbf{r} \end{aligned} \right\} \quad (3)$$

partially decouples the electrons from the phonons and extends the perturbation-theoretical results. The binding energy (in units of $\hbar\omega$) is

$$E_0 = -\alpha + O(\alpha^2) \quad (4)$$

and the effective mass

$$m^* = 1 + \frac{1}{6}\alpha \quad (5)$$

The result confirms the view of Fröhlich, Pelzer and Zienau² that self-trapping does not occur in, for example, the alkali halides ($\alpha \approx 5$ for NaCl).

For very large α ("strong coupling", $\alpha \gtrsim 10$), a variational treatment by Pekar⁴ *does* lead to very large effective masses; he finds

$$E_0 \approx -0.10\alpha^2 \quad (6)$$

$$m^* \approx 230(\alpha/10)^4 \quad (7)$$

However, for intermediate values of α , neither method is very good. A method due to Feynman⁵ provides a "bridge" between the strong- and intermediate-coupling regimes. The electron propagator (for states with no free phonons) is expressed as an integral over all possible paths of the exponential of an action functional. The oscillator coordinates do not appear explicitly, but the action contains a term which represents an interaction of the electron *with itself at earlier times*, through a Coulomb potential:

$$\langle \mathbf{r}_1 t' | \mathbf{r}_2 t'' \rangle = \int \mathcal{D} \mathbf{r}(t) e^S \quad (8)$$

$$S = - \int_{t'}^{t''} \frac{1}{2} \dot{\mathbf{r}}^2 dt + \frac{\alpha}{2} \iint_{t'}^{t''} dt_1 dt_2 \frac{\exp - (t_1 - t_2)}{|\mathbf{r}(t_1) - \mathbf{r}(t_2)|} \quad (9)$$

[For convenience, we use an imaginary time variable.]

Feynman makes the variational ansatz for the action

$$S_0 = - \frac{1}{2} \int_{t'}^{t''} \dot{\mathbf{r}}^2 dt - C \iint_{t'}^{t''} dt_1 dt_2 (\mathbf{r}(t_1) - \mathbf{r}(t_2))^2 \exp - w(t_1 - t_2) \quad (10)$$

where C and w are parameters to be determined. This trial functional is the exact action for a model system in which the electron is coupled to a fictitious particle of mass $M = \frac{4C}{w^3}$ through a

spring constant $K = \frac{4C}{w}$, after the coordinates of the fictitious particle have been eliminated. An upper bound for the energy can be derived:

$$E = \frac{3}{4} \frac{(v - w)^2}{v} - \frac{\alpha}{\sqrt{\pi}} \frac{v}{w} \int_0^\infty dt e^{-t} \left\{ t \left(1 + \frac{v^2 - w^2}{vw^2} \left[\frac{1 - e^{-vt}}{t} \right] \right) \right\}^{-1} \quad (11)$$

where $v^2 = (4C/w) + w^2$. Equation (11) then has to be minimized with respect to w and v . In the limits $\alpha \lesssim 3$, $\alpha \gtrsim 10$, it reduces to the Lee-Low-Pines³ and Pekar⁴ solutions respectively, but for intermediate values E has had to be evaluated numerically⁵. The effective mass and the mobility have also been calculated, and vary smoothly with α .

When α is very large, the radius of the polaron becomes small. A measure of the radius, R , is the amplitude of zero-point oscillation of the two-particle model system,

$$R \sim \left(\frac{v^2 - w^2}{3v} \right)^{-1} \quad (12)$$

For large α , R is found to be $\approx \alpha^{-1}$. Since the unit of length, $\sqrt{\hbar/mw} \sim 10\text{\AA}$, R can become less than the lattice spacing; the continuum model must then break down. In this case (the "small" or "localized" polaron), Fröhlich and Sewell⁷ showed that the appropriate model is a Bloch tight-binding one; the overlap integrals between polaron wave functions localized on adjacent lattice sites are assumed small. The polaron mass is then very large. Conduction at finite temperatures occurs largely by the electron "hopping" from one site to another—the jump frequency is related to the overlap integrals.

Experimental tests of polaron concepts have been limited mainly by the difficulty of measuring the "bare" mass m , and hence of knowing the value of α . However, recent experiments by F. C. Brown⁸ and his collaborators have largely succeeded in "disentangling" the parameters. They have compared the "Hall" mobility μ_H (defined as cR/ρ , where R is the Hall coefficient and ρ the resistivity) with the Ohmic or "drift" mobility μ ; using the theory of Feynman *et al.*,⁸ they are able to find explicit values of m , m^* and α . They have also observed cyclotron resonance of charge carriers in an ionic medium;

this gives an independent estimate of m^* . In principle, one should be able to measure m by looking at a cyclotron resonance *above* the Reststrahl frequency, but this has not yet proved possible.

C. G. KUPER

References

- General reference:* Kuper, C. G., and Whitfield, G. D. Eds., "Polarons and Excitons," Edinburgh, Oliver & Boyd, Ltd., 1963.
- Landau, L. D., *Phys. Z. Sowjetunion*, **3**, 644 (1933).
 - Fröhlich, H., Pelzer, H., and Zienau, S., *Phil. Mag.*, **41**, 221 (1950).
 - Lee, T.-D., Low, F., and Pines, D., *Phys. Rev.* **90**, 297 (1953); Gurari, M., *Phil. Mag.*, **44**, 329 (1953).
 - Pekar, S. I., *Zh. Eksperim. i Teor Fiz.*, **16**, 335, 341 (1946); Allcock, G. R., "Polarons and Excitons," p. 45, 1963.
 - Feynman, R. P., *Phys. Rev.*, **97**, 660 (1955); Schultz, T. D., *Phys. Rev.*, **116**, 526, (1959).
 - Feynman, R. P., Hellwarth, R. W., Iddings, C. K., and Platzman, P. M., *Phys. Rev.*, **127**, 1004 (1962).
 - Fröhlich, H., and Sewell, G. L., *Proc. Phys. Soc.*, **74**, 643 (1959); Sewell, G. L., *Phil. Mag.*, **3**, 1361 (1958). Holstein, T., *Ann. Phys.*, **8**, 343, (1959).
 - Brown, F. C., "Polarons and Excitons," p. 323, 1963. Ascarelli, G., "Polarons and Excitons," p. 357, 1963.

Cross-references: EXCITONS, POLAR MOLECULES.

POLYMER PHYSICS

Polymers are long-chain molecules with molecular weights of from thousands to many millions (generally between 20,000 and 10^7 for materials of practical interest). The molecules may be linear, branched, or cross-linked to give a gel structure.

The size and shape of polymer molecules are generally determined from measurements in dilute solution. Molecular weights are obtained from osmotic pressure (number average molecular weight), light scattering (weight average molecular weight), and intrinsic viscosity measurements. Dissymmetry of light scattered by dilute solutions gives the size of molecules. Extent of chain branching can be estimated from light scattering or solution viscosity measurements by comparing the branched polymer with a linear polymer of the same molecular weight. The degree of cross-linking can be determined from the extent of swelling in a solvent or from the elastic modulus by using the kinetic theory of rubber. Swelling decreases and modulus increases as cross-linking increases. Solution properties are not only sensitive to molecular weight but also to the interaction between the polymer and solvent molecules. Most polymers have a distribution of molecular weights. In addition to the results from fractionations, the width of the distribution can be estimated from the ratio of

weight average to number average molecular weights. This ratio is 1.0 if all the molecules are the same; it is around 2 for most polymers, but may be much greater for some highly branched polymers.

The molecular structure of polymers is unusually complex since the molecules can assume many conformations; more than one type of monomeric unit can make up the chains to give an infinite variety of distribution of sequence lengths, or the monomeric units can be arranged in different types of stereoregularity—isotactic, syndiotactic, or atactic forms. Nuclear magnetic resonance and infrared spectroscopy are especially powerful techniques for studying the structure of polymers. For polymers capable of crystallizing, x-ray diffraction is another useful tool.

The most important quantity determining the mechanical and many other physical properties of polymers is the glass transition temperature. If the glass transition temperature, T_g , is below ambient temperature, the molecules have extensive freedom of movement, so the material is either a viscous liquid or a rubbery material with a low elastic modulus. If T_g is above ambient temperature, the movement of the molecules is frozen-in, so that the polymer is a rigid solid with a high elastic modulus of the order of 10^{10} dynes/cm². The glass transition temperature is not sharply defined but depends to some extent on the time scale of the experiment—the faster the experiment, the higher is the apparent T_g . Glass transitions may be measured by many techniques such as where breaks occur in the slope of volume or refractive index vs temperature curves or by the rapid change in elastic modulus with temperature in the transition region. The position of T_g on the temperature scale is largely due to the stiffness of the polymer chains. Flexible molecules such as polybutadiene and silicone rubbers have low T_g , while stiff molecules such as polystyrene and polymethyl methacrylate have high transition temperatures. Cohesive energy density or polarity is another important factor in determining T_g . Symmetry plays a secondary role. Glass transitions can be regulated by copolymerization or by addition of a plasticizer which lowers T_g .

Many polymers including polyethylene and isotactic polypropylene are semicrystalline. In their bulk behavior such polymers behave as though they are a mixture of amorphous and crystalline materials, but the exact nature of the crystalline state is not yet clearly defined for such materials. In the crystal lattice, some types of polymer chains assume a zig-zag conformation while others crystallize in the form of helices. Single crystals of some polymers have been grown from dilute solutions. In these single crystals the chains are perpendicular to the faces making up the thin lamellar crystals, so that each polymer chain must fold back on itself several times. There is some morphological evidence, based on electron microscopy studies, that even in the bulk polymer cooled down from the melt there

is extensive chain folding in the crystalline phase. In terms of a two-phase model, the degree of crystallinity may be determined by x-ray, density or heat capacity measurements. Different techniques generally give similar but not identical values for the degree of crystallinity; typical values vary from 40 per cent crystallinity for low-density polyethylene to 85 per cent for high-density polyethylene. Crystallinity is greatly affected by chain perfection. Copolymerization and branching greatly reduce crystallinity. Highly stereoregular polymers such as isotactic polystyrene tend to be crystalline, while the random atactic polymers are noncrystalline. The melting point also depends upon chain perfection—the greater the degree of imperfection, the lower is the melting point.

Many polymers of commercial importance are not linear polymers consisting of a single type of monomeric unit. Copolymers contain two or more kinds of monomers. If one type of monomer makes up the backbone and another type side chains, the polymers are called graft polymers. If two polymers are mechanically mixed together, the mixture is called a polyblend. Most, but not all, polyblends are two-phase systems.

Even at higher temperatures where linear polymers are liquid, they tend to be very viscous. The melt viscosity is especially high if the molecular weight is above a critical value where chain entanglements can occur. At molecular weights above which entanglements occur, the melt viscosity at low rates of shear depends approximately on (molecular weight)³⁻⁴. These viscous polymer melts are also more or less elastic in nature and behave somewhat like rubber. If the molecules are cross-linked to one another, the elastic behavior becomes dominant and true vulcanized rubbers result. Polymer melts are generally non-Newtonian, and the properties are very dependent upon the rate of shear. Molecular theories have been developed by Rouse, Zimm, and Bueche which explain quite well many of the rheological properties of melts and solutions.

The usefulness of polymers depends primarily upon their mechanical properties. The elastic modulus of rubbers is explained quite satisfactorily by the kinetic theory of rubber elasticity. No satisfactory theory is yet available for rigid polymers. Dynamic mechanical measurements using oscillating stresses or strains have been especially useful in relating mechanical properties to molecular structure; such tests are generally made to measure the elastic modulus and mechanical damping over a wide range of frequencies and temperatures. Generally the effects of temperature and frequency (or time) can be made equivalent by a superposition treatment such as developed by Williams, Landel, and Ferry. The phenomenological theory of viscoelasticity has developed to the stage where it is possible to interconvert data from one type of test (say dynamic mechanical) to other types of tests such as creep, stress relaxation and, to a lesser extent, stress-strain data. On heating a rigid organic polymer, the modulus drops from about 10^{10}

dynes/cm² to a low value of about 10^7 dynes/cm² in a small temperature interval near the glass transition temperature, unless the polymer is highly cross-linked or crystalline. Both crystallinity and cross-linking can greatly increase the modulus above T_g , but they have little effect on the modulus below T_g . Polymers are unique in that some of them can be elongated over 1000 per cent before they break.

The dielectric constant and power factor or electrical loss depend upon the number and type of dipoles. At low temperatures (or high frequencies) where the dipoles are frozen-in, the dielectric constant and electrical loss are both low. At high temperatures (or low frequencies) where the molecules have high mobility, the dielectric constant is high while the electrical loss is often low. At intermediate temperatures or frequencies where the main relaxation times for dipolar motion are approximately the same as the applied electrical frequency, the electrical loss goes through a pronounced maximum, and the dielectric constant changes rapidly with either frequency or temperature. Since the mobility of the chain backbone is related to the ease with which dipoles can move, there is often a good correlation between electrical and mechanical properties. Impurities in polymers can be very detrimental to good electrical properties, especially at high temperatures where conductivity can be relatively high.

Most pure amorphous polymers are transparent. Crystallinity often makes a material milky or white in appearance. Long chain molecules can be oriented by stretching in the molten state or by cold-drawing in some cases. Such oriented materials are generally highly birefringent, since the polarizability along the chain is usually quite different from the polarizability perpendicular to the chain. The mechanical properties of oriented polymers are also highly anisotropic. For instance, the modulus and tensile strength are generally much greater parallel to the chain axis than perpendicular to it.

Polymers have very high coefficients of thermal expansion compared to most rigid materials. The coefficient of expansion shows a distinct break at the glass transition temperature—the coefficient being greater above T_g . Most polymers would be classed as thermal insulators rather than as thermal conductors.

LAWRENCE E. NIELSEN

References

- Bueche, F., "Physical Properties of Polymers," New York, Interscience Publishers, 1962.
- Nielsen, L., "Mechanical Properties of Polymers," New York, Reinhold Publishing Corp., 1962.

Cross-references: DIELECTRIC THEORY, LIGHT SCATTERING, MOLECULAR WEIGHT, MOLECULES AND MOLECULAR STRUCTURE, OSMOSIS, VISCOELASTICITY, VISCOSITY.

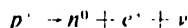
POSITIVE RAYS. *See* ISOTOPE

POSITRON

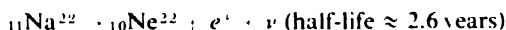
The positron is one of many fundamental bits of matter. Its rest mass (9.109×10^{-31} kg) is the same as the mass of the electron, and its charge ($+1.602 \times 10^{-19}$ coulomb) is the same magnitude but opposite in sign to that of the electron. The positron and electron are antiparticles for each other. The positron has spin 1/2 and is described by Fermi-Dirac statistics as is the electron (see ELECTRON).

The positron was discovered in 1932 by C. D. Anderson at the California Institute of Technology while doing cloud chamber experiments on cosmic rays. The cloud chamber tracks of some particles were observed to curve in such a direction in a magnetic field that the charge had to be positive. In all other respects, the tracks resembled those of high-energy electrons. The discovery of the positron was in accord with the theoretical work of Dirac on the negative energy states of electrons. These negative energy states were interpreted as predicting the existence of a positively charged particle.

Positrons can be produced by either nuclear decay or the transformation of the energy of a gamma ray into an electron-positron pair. In nuclei which are proton-rich, a mode of decay which permits a reduction in the number of protons with a small expenditure of energy is positron emission. The reaction taking place during decay is



where p^+ represents the PROTON, n^0 the NEUTRON, e^+ the POSITRON, and ν a massless, chargeless entity called a NEUTRINO. The positron and neutrino are emitted from the nucleus while the neutron remains bound within the nucleus. An example of such a nuclear decay is



This particular decay provides a practical, usable source of positrons for experimental purposes.

The process of pair production occurs when a high-energy gamma ray interacts in the electromagnetic field of a nucleus to create a pair of particles—a positron and an electron. Pair production is an excellent example of the fact that the rest mass of a particle represents a fixed amount of energy. Since the rest energy ($E_{\text{rest}} = m_{\text{rest}}c^2$) of the positron plus electron is 1.022 MeV, this energy is the gamma energy threshold and no pair production can take place for lower-energy gammas. In general, the cross section for pair production increases with increasing gamma energy and also with increasing Z number of the nucleus in whose electromagnetic field the interaction takes place.

The positron is a stable particle (i.e., it does not decay itself), but when it is combined with its antiparticle, the electron, the two annihilate each other and the total energy of the particles appears in the form of gamma rays. Before

annihilation with an electron, most positrons come to thermal equilibrium with their surroundings. In the process of losing energy and becoming thermalized, a high-energy positron interacts with its surroundings in almost the same way as does the electron. Thus for positrons, curves of distance traversed in a medium as a function of initial particle energy are almost identical with those of electrons.

It is energetically possible for a positron and an electron to form a bound system similar to the hydrogen atom, with the positron taking the place of the proton. This bound system has been given the name "positronium." The energy levels of positronium are about one-half those of the hydrogen atom since the reduced mass of positronium is about one-half that of the hydrogen atom. This also causes the radius of the positronium system to be about twice that of the hydrogen atom. Thus positronium is a bound system with a radius of 1.06 Å and a ground state binding energy of 6.8 eV. As mentioned previously, the positron has an intrinsic magnetic moment and an intrinsic angular momentum, or spin. In positronium, the spins of the electron and positron can be oriented so they are either parallel or antiparallel. These two states, called ortho-positronium and para-positronium respectively, have very different annihilation characteristics. Most positrons entering a medium do not form positronium, but the general annihilation characteristics show the same dependence on orientation of the spins, regardless of whether the annihilation occurs in a collision or from the bound state.

It is possible for a positron-electron system to annihilate with the emission of one, two, three, or more, gamma rays. However, not all processes are equally probable. One-gamma annihilation requires another particle to participate to conserve momentum. This process is a very infrequent type decay. The most probable decay is by the emission of two gamma rays, directed in opposite directions, with each possessing about one-half the energy of the system. The presence of these 0.511-MeV ($=m_e c^2$) gamma rays is always found when positrons are present. Whether annihilation is to be by the emission of one, two or three gammas depends on the orientation of the spins of the positron and electron. Conservation of angular momentum requires that the decay be by two-gamma emission if the spins are antiparallel, and by three-gamma (or one-gamma) emission if the spins are parallel.

If formed in free space, positronium exhibits two characteristic lifetimes against self-annihilation. These are $\tau_1 = 1.25 \times 10^{-10}$ second for the antiparallel spin case (also called the singlet state) and $\tau_3 = 1.39 \times 10^{-7}$ second for the parallel spin case (called the triplet state). Another lifetime, characteristic of the physical surroundings of the positron, is found when positronium is formed in certain condensed materials. This lifetime (known as τ_2) is longer than the singlet free space lifetime τ_1 , but is much shorter than

the triplet free space lifetime τ_3 . In general, this τ_3 lifetime is a measure of the rate of "pickoff" of atomic electrons with antiparallel spins by the positrons in triplet positronium. In this process, the positron enters the material and forms triplet positronium with an electron, but then annihilates with an electron belonging to one of the surrounding atoms, whose spin is oriented opposite to that of the positron. That the probability of "pickoff" and the subsequent two-gamma annihilation depend on the properties of the surroundings is to be expected. Indeed, the τ_3 lifetime is a function of the material, the temperature, the density, the degree of crystallinity, the phase, etc.

The angular correlation of the annihilation gammas has been measured for both the two-gamma and three-gamma cases. In three-gamma annihilation, the gammas are coplanar as predicted, and azimuthally correlated such that their energies and directions are consistent with the conservation laws of energy and momentum. Two-gamma annihilation studies have been extensive, the results showing that the two gammas are emitted within a few milliradians of 180° from each other. The departures from the expected angular distributions can be used to gain information on the momentum distribution of the electrons with which the positrons annihilate. Thus positron annihilation becomes a tool to learn more concerning the internal structure of materials.

B. CLARK GROSECLOSE

References

1. Berko, S., and Hereford, F. L., *Rev. Mod. Phys.*, **28**, 299 (1956).
2. Ferrell, R. A., *Rev. Mod. Phys.*, **28**, 308 (1956).
3. Green, J., and Lee, J., "Positronium Chemistry," New York, Academic Press, 1964.
4. Wallace, P. R., *Solid State Phys.*, **10**, 1 (1960).

Cross-references: ATOMIC PHYSICS, ELECTRON, ELEMENTARY PARTICLES, NEUTRINO, NEUTRON, NUCLEAR STRUCTURE, PROTON.

POTENTIAL

The earliest concept of potential was developed for electric phenomena by Simeon Poisson, George Green and others in the period from about 1813 to 1827. It was realized that energy could be released from electrically charged bodies, and the concept of potential was developed as one characteristic of such a charged system that measured the ability of the system to release this energy. The energy stored in the system was and is called *potential energy*, as contrasted to dynamic or kinetic energy, hence the acceptance of the term *potential*. This concept is a scalar quantity as contrasted to a vector quantity.

Although electric potential is the most usual form of potential, other forms have also been defined to serve particular needs. These include

the scalar magnetic potential and the vector magnetic potential to be discussed later.

Closely associated with the concept of electric potential was the concept of *charge* that had been formulated by Charles Augustin de Coulomb and others over about a 100-year period starting about 1737. The electric scalar potential of a macroscopic system is defined as the electric energy of the system divided by the electric charge. This simple definition presumes a basic two-conductor, statically charged system. The mathematical expression for the definition of the potential Φ is

$$\Phi = \frac{W}{Q}$$

where W is the energy of the system and Q is the charge (see list of units at end of article).

Modern electric systems employ the concept of potential for much more sophisticated forms through the use of summations of charge effects either in discrete or in distributed forms. The complete expression for electric potential caused by the accumulation of a number of concentrated charges Q_k , surface charge density σ over a surface S , and volume charge density ρ in a volume V , all being located in a dielectric media of uniform permittivity ϵ , is

$$\Phi = \sum_{k=1}^n \frac{Q_k}{4\pi\epsilon r} + \frac{1}{4\pi\epsilon} \int_S \frac{\sigma ds}{r} + \frac{1}{4\pi\epsilon} \int_V \frac{\rho dv}{r}$$

where r measures the magnitude of the distance from the point at which the electric potential is evaluated to each charged particle or element of charge.

The difference in electric potential from one point a to another point b in a static electric field is related to a vector function called the electric field intensity \mathcal{E} (a characteristic of the space related to the negative of the gradient of the electric potential) through the line integral relationship

$$\Phi_{ab} = - \int_a^b \mathcal{E} \cdot d\mathbf{l},$$

where \mathbf{l} is a vector direction measured along the path in the direction from the point a to the point b , and the scalar product between \mathcal{E} and $d\mathbf{l}$ is denoted by the dot or scalar product notation.

The inverse form of this integral expression is the gradient form for the static electric field intensity, namely

$$\mathcal{E} = - \nabla \Phi$$

where the ∇ symbol is a vector differential space operator which, when applied to the scalar electric potential Φ , yields the electric potential gradient.

An extension of spatial derivative operations yields another function of the scalar electric potential Φ , known as the scalar Laplacian, as $\nabla^2 \Phi$. This term, when equated to the negative of the volume space charge density ρ divided by the permittivity of the space, is known as Poisson's equation. Thus,

$$\nabla^2 \Phi = -\rho/\epsilon$$

If the electric volume charge density is zero, the above relationship becomes

$$\nabla^2 \Phi = 0$$

and is known as Laplace's equation.

Electric charges in motion produce magnetic effects, which result in a magnetic field intensity vector \mathbf{H} , somewhat analogous to the corresponding vector \mathcal{E} in electric field phenomena. The line integral relationship between this vector \mathbf{H} and the vector direction measured along a path in space from point a to b establishes an analogous scalar magnetic potential Ψ_{ab} as

$$\Psi_{ab} = - \int_a^b \mathbf{H} \cdot d\mathbf{l}$$

This function is useful in the evaluation of magnetic field geometric relations for locations in space where the electric current density is zero.

For regions in magnetic fields where current densities exist, a designation of a magnetic potential relation can only be made through a function called the *vector magnetic potential*. This vector is defined in a form similar to the expression for scalar electric potential caused by a volume distribution of electric charge density as given earlier. Thus, the vector magnetic potential \mathbf{A} at a point in space caused by a distribution of current density \mathbf{J} over a volume V is

$$\mathbf{A} = \frac{\mu}{4\pi} \int_V \frac{\mathbf{J} dv}{r}$$

where μ is the permeability of space and r is the magnitude of the distance from the point at which \mathbf{A} is evaluated to each element of current density of the system. The point under consideration can be a point within the region of current density.

The vector magnetic flux density \mathbf{B} in such a space is related to the vector magnetic potential \mathbf{A} through a spatial differential function called the curl and symbolized by the operational form as

$$\mathbf{B} = \nabla \times \mathbf{A}$$

An extension of spatial derivative operations upon the vector magnetic potential also yields another function of this potential known as the vector Laplacian, $\nabla^2 \mathbf{A}$. This term is related to the current density \mathbf{J} at a point in the field as

$$\nabla^2 \mathbf{A} = -\mu \mathbf{J}$$

This expression is analogous to the similar scalar Laplacian of the electric potential which was related to the electric charge density at the point of evaluation.

For systems in which the charges are moving in such a manner that the vector magnetic potential is not constant with respect to time, the elementary form of the relationship between the electric field intensity \mathcal{E} and the scalar electric potential Φ must be modified to include a

function of the vector magnetic potential, namely

$$-\nabla \Phi - \frac{\partial \mathbf{A}}{\partial t}$$

This is the general expression that is valid at a point in space for all conditions. As action at a distance is considered, the elementary forms for the evaluation of the scalar electric potential Φ and of the vector magnetic potential \mathbf{A} , however, must recognize the time delay in action with respect to the causes.

If the scalar electric potential at a point P is to be evaluated at some time t , the *retarded potential* must consider the finite velocity of propagation that occurs in the path length r between the point P and the location of the charge that causes the electric potential.

Depending upon the nature of the charge, the retarded potential expression can be expressed in terms of the retarded time $t - r/c$, where c is the velocity of the propagated effect in free space. For the case in which the charge is distributed over a volume, the expression becomes

$$\Phi_{P,t} = \frac{1}{4\pi\epsilon} \int_V \frac{[\rho]_{t-r/c} dv}{r}$$

The symbol $[\rho]_{t-r/c}$ indicates that the charge density at the source of the field is that evaluated at an earlier time $t - r/c$. If the charges are discrete or distributed over surfaces, the corresponding forms of the potential function for these geometries would be used.

In a similar manner, the retarded vector magnetic potential can be expressed for a volume distribution of current density as

$$\mathbf{A}_{P,t} = \frac{\mu}{4\pi} \int_V \frac{[\mathbf{J}]_{t-r/c} dv}{r}$$

All preceding relations are expressed in a form that results if a rationalized system of units is used. The internationally accepted units that conform with that preceding relationships are:

Entity	Symbol	Unit
Energy	W	joule
Potential (electric)	Φ	volt
Charge	Q	coulomb
Length	r, l	meter
Area	S	meter ²
Volume	V	meter ³
Surface charge density	σ	coulomb/meter ²
Volume charge density	ρ	coulomb/meter ³
Permittivity	ϵ	farad/meter
Electric field intensity	\mathcal{E}	volt/meter
Magnetic field intensity	\mathbf{H}	ampere/meter
Scalar magnetic potential	Ψ	ampere
Vector magnetic potential	\mathbf{A}	weber/meter
Permeability	μ	weber/meter-ampere
Magnetic flux density	\mathbf{B}	tesla
Current density	\mathbf{J}	ampere/meter ²
Time	t	second
Velocity	c	meter/second

WARREN B. BOAST

References

- Boast, W. B., "Vector Fields," New York, Harper and Row, 1964.
- Plonsey, R., and Collin, R. E., "Principles and Applications of Electromagnetic Fields," New York, McGraw-Hill Book Co., 1961.
- Javid, M., and Brown, P. M., "Field Analysis and Electromagnetics," New York, McGraw-Hill Book Co., 1963.
- Moon, P., and Spencer, D. E., "Foundations of Electrodynamics," Princeton, N. J., D. Van Nostrand Co., 1960.

Cross-references: ELECTRICITY, ELECTROMAGNETIC THEORY, STATIC ELECTRICITY.

POWER. See WORK POWER AND ENERGY

PRESSURE. See FLUID STATICS

PRESSURE, VERY HIGH

For the purposes of this article "very high pressure" is defined as the range above 50 kilobars. Discussion is confined to apparatus and experiments capable of performance in this range. Thus, no discussion of the vast array of sophisticated measurements in the 12 to 20 kilobar range is included.

The increasing interest in research at very high pressure has been stimulated in part by the synthesis of diamond at General Electric, in part by an expansion of experimental GEOPHYSICS, and in large part by an increased appreciation of the importance of relatively large variations of interatomic distance in our understanding of the electronic structure of solids.

It is convenient to divide the available types of equipment into those capable of chemical synthesis on a reasonable scale and those primarily useful for physical measurements.

Among the former group, the most straightforward is the piston and cylinder device brought to its highest development by Kennedy and his colleagues at UCLA. With appropriate support it is capable of 70 kilobars. In this apparatus the pressure determination is more direct and probably more accurate than in the others.

The most generally useful apparatus for relatively large scale work at higher pressures is the General Electric "belt." As originally designed by Hall, pressures of perhaps 140 kilobars were possible. Later modifications by Bundy have extended this range by another 40 to 50 kilobars. The essential feature of the device is support for the pistons which increases with increasing pressure.

A third "large volume" apparatus in general use is the tetrahedral press first developed by Hall and modified by Hutton. In its smaller scale versions it appears capable of 100 kilobars or more, but the larger scale-ups are somewhat more limited.

All three of the above types of equipment can be operated at elevated temperatures, to 1000°C or beyond, although the accuracy of both pressure and temperature measurements is limited under these extreme conditions.

By far the most common "small scale" equipment used for physical measurements at high pressure is the tapered anvil design originated by P. W. Bridgman. It has the great advantage of relative cheapness of construction and ease of loading. While pressure gradients are a problem, they have not limited its broad application. Originally developed for electrical resistance measurements, it has since been applied to x-ray scattering, to Mössbauer studies, and (with diamond anvils) to optical absorption studies. While there have been a number of estimates of the pressure range of Bridgman anvils, the probable upper limit is near 150 kilobars.

A modification of the tapered anvil apparatus has given the highest pressures yet obtained statically. This involves very small flats, work hardened, sintered, tungsten carbide pistons, and support on the taper which increases with increasing pressure, so that the net effect is that measurements are made in a cell surrounded by rings of material at continuously decreasing pressure. With modifications of this apparatus, optical absorption measurements have been made to 160 to 170 kilobars, electrical resistance and x-ray diffraction studies to over 500 kilobars, and Mössbauer studies to beyond 250 kilobars. The apparatus has been operated from 77 to 650°K, but increasing temperature above 300 K limits the pressure range rapidly.

Finally, mention should be made of shock velocity measurements. With this technique p - v measurements have been made at pressures of 2000 kilobars both at Los Alamos and by Russian workers. Measurements have largely been confined to p - v relations to date, but other types are apparently possible. The correction from adiabatic experiments to isothermal data limits the accuracy for very compressible materials.

The problem of calibration is a complex one, and only an outline can be given here. The most usual method has been to observe discontinuities in electrical resistance due to first-order phase changes in certain metals. A serious problem is that these materials exhibit varying degrees of metastability which are not independent of the apparatus. The calibration problem at elevated or reduced temperature is still more serious. Points most frequently used at 300°K are transitions in bismuth at 25 and 87 kilobars, in thallium at 37 kilobars, in barium at 59 and 140 kilobars, in iron at 130 kilobars, and in lead at approximately 160 kilobars. Beyond this point, the difficulties increase. By means of x-ray measurements on systems where shock wave data are available, consistency between shock and static measurements has been established to over 500 kilobars. Internal consistency among some shock measurements has also been shown by x-rays on mixed powders. One certain conclusion is that extensive extrapolation of linear calibrations established at low pressure invariably predicts pressures much higher than are obtainable.

In a brief article such as this it would be impractical to review in any detail the experimental results, but an outline of major features is feasible.

A great deal of careful effort has gone into the establishment of melting curves and phase boundaries between two solid phases. Items of particular interest in the first category include maxima in the melting points of cesium and barium discussed by Kennedy and co-workers. High pressure phase transitions to a metallic state have been discovered in silicon, germanium, gallium arsenide, gallium antimonide, aluminum antimonide, indium arsenide, indium phosphide, indium antimonide, zinc sulfide, zinc selenide, zinc telluride, and other III-V and III-IV compounds. Recent x-ray studies indicate that the high pressure phases of silicon and germanium have the white tin structure while a number of the III-V compounds have a closely related atomic arrangement. On the other hand, ZnSe and ZnTe appear to adopt the face-centered-cubic NaCl structure at high pressure. In the case of InSb and a few related compounds it has been possible to quench in the high pressure phase at one atmosphere and to show that these are indeed SUPERCONDUCTORS below about 3°K. (See CRYSTALS AND CRYSTALLOGRAPHY and SEMICONDUCTIVITY.)

The General Electric group have shown that after compression into the metallic phase and release of pressure, silicon and germanium retain rather a complex cubic or tetragonal arrangement which is neither white tin nor diamond. It has not yet been shown whether these intermediate phases have any definite range of true stability. Moderate heating at one atmosphere suffices to return these materials to the diamond structure.

The pressure-volume measurements by Swenson and his co-workers at Iowa State on alkali metals and solidified rare gases at liquid helium temperature have given an important impetus to our understanding of interatomic forces in relatively simple systems.

Shock wave studies have provided p - v data on nineteen metals to 2000 kilobars and a number of others to 500 kilobars, as well as compressibilities of a number of alkali halides to the 300 to 400 kilobar region.

X-ray measurements have shown that the high pressure phase of iron is hexagonal close packed with a very anisotropic compressibility (c/a is 1.64 at 150 kilobars and 1.58 at 400 kilobars). Recent measurements on lattice parameters of magnesium reveal a rather complex behavior for the c/a ratio which can be related to interaction between the Fermi surface and the Brillouin zone boundaries.

A very interesting high pressure phenomenon is the electronic transition. Cesium exhibits a cusp in the resistance at 41 kilobars and a second maximum above 200 kilobars. Rubidium has a distinct maximum in its resistance above 400 kilobars, while potassium shows a very sharp rise in resistance (by a factor over 50) in 600 kilobars. These phenomena have been associated with the promotion of an s electron to the empty d shell which is at higher energy in the free atom and the low pressure solid. The increase

in pressure at which this takes place going from cesium to potassium can be associated with the greater energy difference between $3d-4s$ as compared with $4d-5s$ and $5d-6s$ shells.

Cerium has a transition at five kilobars which is associated with $5d-4f$ electronic mixing, and many of the RARE EARTH metals exhibit unusual resistance behavior at high pressure which can almost certainly be related to mixing between $4d$, $5f$, and other nearby levels. Strontium and calcium among the alkaline earth metals exhibit transitions to a semiconducting or semimetallic state at high pressure and then ultimately become metallic again.

In insulators and SEMICONDUCTORS, one of the more significant studies involves the shift of the optical absorption edge (gap between the conduction and valence band) with pressure and the accompanying change in electrical resistance. A number of features concerning the band structure of silicon, germanium, and related compounds have been revealed by these measurements.

The continuous approach to the metallic state is best illustrated by studies on iodine where a very close agreement is shown between the activation energy for electrical conduction and half the optical energy gap. Between 160 to 240 kilobars the resistance is metallic in one direction but not in the other, much like graphite. At higher pressures the behavior is distinctly metallic.

High pressure optical studies on the alkali halides have been important in elucidating the electronic structure in the neighborhood of imperfections. The local compressibility of the F center is about twice the bulk compressibility at low pressure and decreases to perhaps 1.2 to 1.4 times the bulk compressibility above 100 kilobars. This is quite consistent with the picture of a vacancy containing a trapped electron.

A well-known phosphor involves a thallous ion as a substitutional impurity in an alkali halide lattice. The pressure dependence of the optical absorption due to this impurity shows that the transition involved is closely bound to the thallous ion in the halides, but is more spread out in the iodides.

Crystal field theory has been a very important first-order picture of the ENERGY LEVELS of transition metal ions in crystals and complexes. The effect of pressure on the transition energy shows very clearly the applicability and limitations of the theory. In rigid ionic networks such as Al_2O_3 , the energy increases as the inverse fifth power of the interatomic distance, as required by simple crystal field theory. In most cases, however, it is clear that the point "charge" assumption of crystal field theory is inadequate for quantitative calculations, and increasing pressure results in spreading out of the electron cloud and decreased interelectronic repulsion among the $3d$ electrons.

It is not hard to see that there is scarcely an area of solid-state physics where high pressure measurements are not capable of providing a significant test of theory.

A number of good references are available. Bridgman's¹ monograph remains a classic. The book edited by Wentorf² and Swenson's³ review paper give excellent discussions of techniques as of 1961, although much has developed since. Recent books edited by Warshauer and Paul⁴ and by Bradley⁵ cover many applications. The former is of especial interest to solid state physicists. A paper in *Advances in Chemical Physics*⁶ reviews some studies on electronic structure.

H. G. DRICKAMER

References

1. Bridgman, P. W., "Physics of High Pressure," Second edition, New York, Bell and Company, 1949.
2. Wentorf, R. H., Ed., "Modern Very High Pressure Techniques," London, Butterworths, 1962.
3. Swenson, C. A., in Seitz, F., and Turnbull, D., Eds., "Solid State Physics," Vol. II, New York, Academic Press, 1960.
4. Paul, W., and Warshauer, D. M., Eds., "Solids Under Pressure," New York, McGraw-Hill, 1963.
5. Bradley, R. S., Ed., "Physics and Chemistry of High Pressure," New York, Academic Press, 1963.
6. Prigogine, R., Ed., *Advan. Chem. Phys.*, 4 (1963).

PROPAGATION OF ELECTROMAGNETIC WAVES

The discussion of propagation phenomena in general media requires a rather elaborate and sophisticated mathematical formalism. For a detailed treatment of those disturbances that fall into a category of wave phenomena, an ELECTROMAGNETIC THEORY of characteristics whose origin lies in the subject of partial differential equations has long been available.

Two viewpoints are of interest for an understanding of the mechanism of propagation of an electromagnetic wave: the macroscopic phenomenological theory of Maxwell and the microscopic theory of Lorentz. The Maxwell theory avoids the explanation of what happens to the individual charged particles that constitute the medium in which the phenomena are taking place; the influence of the medium is accounted for by means of factors called constitutive parameters. The effect of the electric field of equal numbers of oppositely charged particles that make up the medium is accounted for by the permittivity ϵ . The effects of the motions of the charges comprising the medium on the fields are taken into account in the following manner: the motion of the unbound charges is subsumed under the conduction current by means of a conductivity σ ; the translational motion of the bound charges is subsumed in the electric-displacement field by means of the permittivity ϵ ; and the rotational motion of the constituent charges is subsumed in the magnetic-induction field by means of the permeability μ .

This representation of the dynamical behavior of the constituent charges by means of the parameters ϵ , μ , and σ has proved to be of great

practical importance, since one can separate the study of the macroscopic behavior of the fields from that of the macroscopic behavior of the constituents of the medium in which the fields are present. In many media these constitutive parameters can be determined by empirical means much more easily than by theoretical considerations. For such media the macroscopic theory is clearly most suitable.

In an ionized gas (in contrast to media made up of neutral molecules), it is difficult to measure the constitutive parameters. The difficulty arises from the fact that the space charges present and the boundary condition imposed by the measuring apparatus affect each other significantly. To determine these parameters a knowledge of the dynamical behavior of the microscopic constituents is required, which leads us to the second point of view, the exhaustive microscopic theory of Lorentz. This theory attempts to describe all electrical phenomena in terms of the elementary positive and negative charges comprising the medium. The theory dispenses with the concept of the material medium and considers only the ensemble of negative and positive charges (which actually constitute the medium) in free space. More specifically, the theory postulates that in all permeable bodies there exists a large number of charged particles of very small size, which are separated from each other by free space. Conducting bodies are imagined to be constituted of a large number of free particles capable of being moved through the body under the action of an electric force. Nonconducting or weakly conducting bodies are considered to be made up of particles bound to their positions of equilibrium by an elastic force. Even though they are displaced from their equilibrium positions, this displacement is not very large (small oscillations). It is also postulated that the medium has no net charge, so that the positive and negative charges balance exactly. When particles are displaced from their equilibrium positions, the medium becomes polarized.

The Lorentz theory further assumes that the free-space displacement current exists not only in the empty space between the particles but also within the particles themselves. The action of the material medium participates in this theory if we consider the motions of the charged particles under the influence of the electromagnetic forces as a fundamental concept. If each particle has a charge q and a mass m , then under the action of the electric force the particle is displaced from its equilibrium position, and at time t has a velocity of magnitude v . However, the moving charge produces a current qv , and if there are N such particles per unit volume, they give rise to a convection current density Nqv (by v we understand the time-average value of the charge velocity). The average convection current density can be written as $Nqv = \rho v = J$ regardless of whether the charges are free or bound.

For varying fields, it is not clear *a priori* whether this total current is a conduction current proportional to the electric intensity and in

phase with it, or a displacement current proportional to $\dot{\mathbf{E}}$ and out of phase with the electric intensity \mathbf{E} by $\pi/2$. Consequently, the total current density must be written as the sum of the free-space displacement current and the material convection current.

Let us summarize the two views: the Maxwell theory describes the phenomena in terms of the equations

$$\nabla \times \mathbf{E} + \dot{\mathbf{B}} = 0 \quad (1)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \dot{\mathbf{D}} \quad (2)$$

$$\nabla \cdot \mathbf{D} = \rho \quad (3)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (4)$$

supplemented by the relations

$$\mathbf{B} = \mu \mathbf{H} \quad (5)$$

$$\mathbf{D} = \epsilon \mathbf{E} \quad (6)$$

$$\mathbf{J} = \sigma \mathbf{E} \quad (7)$$

The Lorentz microscopic theory describes the electromagnetic phenomena by the same set of Maxwell's equations as given above, but since there is no medium in this point of view, the relations (eqs. 5, 6 and 7) must be given in vacuo,

$$\mu = \mu_0 \quad \epsilon = \epsilon_0 \quad (8)$$

and the current density is given directly in terms of the moving charges:

$$\mathbf{J} = \sum_k q_k \mathbf{v}_k / \text{volume of region containing charges} \quad (9)$$

Maxwell's equations relate the fields to the charges and their motions, but since these motions are not known, it is necessary to supplement them with the dynamical equations of motion for the charges. These equations have the form, for each particle q_k ,

$$\frac{d(m_k \mathbf{v}_k)}{dt} = q_k (\mathbf{E} + \mathbf{v}_k \times \mathbf{B}) \quad (10)$$

where \mathbf{B} is the *total* magnetic induction field evaluated at the position \mathbf{r}_k of the k th particle and \mathbf{E} is the *total* electric field evaluated at \mathbf{r}_k . These fields can consist in part of external fields and in part of fields due to the q_k 's themselves.

Equations (1) through (10) must be solved simultaneously for the dynamical behavior of the charges, and from this behavior we can derive the constitutive parameters for the macroscopic model of the medium.

In the Lorentz theory, the motion of the electrical charges is described in terms of the polarization vector (dipole moment per unit volume) rather than in terms of the convection current. In this case the polarization vector $\mathbf{P} = Nq\mathbf{r}$, \mathbf{r} being the average displacement of the charged particles. Consequently, $\mathbf{J} = \dot{\mathbf{P}}$.

In order to discuss propagation in PLASMAS, the relationship between the polarizations (or

the current density) and the electric intensity must be added to Maxwell's equations; this completed set we call the Maxwell-Lorentz equations.

In the ionized gas, the electrons and ions are detached completely from their parent molecules; no elastic forces bind them as in the case of a solid body such as a crystal; the particles have no free period of oscillation of their own. But under the influence of an applied force, they are disturbed by collisions with the neutral molecules. In the collision process some of the energy of motion of the electrons and ions is converted into energy of random motion of particles, i.e., heat energy. The charged particles are therefore losing energy continually, and this loss can be represented as a resistance to their motions. This simple model of a resistive mechanism is incorporated into the equations of motion of the charged particles as a damping force.

The electron theory of Lorentz gives some idea of the mechanism by which the charged particles alter the phase velocity of a propagating electromagnetic wave field. Consider a semi-infinite medium separated by a plane interface from the free space, and imagine an electromagnetic wave field propagating in free space impinging upon the medium. The electromagnetic wave traveling through the free space between the particles excites the charged particles and causes them to oscillate so that they essentially become small dipole oscillators, the wave field governing their phase of oscillation.

At each point either in or external to the region in which the particles are contained, the secondary waves radiated by the oscillators interfere with the original wave field and with the other radiated waves; the sum of all the waves is a resultant field at the point. Moreover, it is the resultant field that acts on the particles at the given point which causes them to vibrate. The total intensity at any point is the vector sum of the intensity of the original wave field and the resultant intensity due to all the oscillating particles. According to this picture, in general, the emerging wave has its phase velocity altered. Reflective waves are treated in a similar manner.

Plane Wave Fields: Dispersion Relations. Consider the kind of waves that an unbounded homogeneous medium can support, in the presence of a uniform (in space-time) externally applied magnetic field. Although this is not the most general situation, a great many of the salient features may be culled from it. The Maxwell-Lorentz equations, together with the postulated magnetic field, permit us to examine the existence of simple plane wave fields, i.e., field quantities proportional to $\exp(-j\mathbf{k} \cdot \mathbf{r})$ where \mathbf{k} is the wave vector whose components are in general, complex. For fields with harmonic time dependence, the above assumptions lead to the condition for propagation which takes the form

$$\Delta(\mathbf{n}, \omega, \dots) \mathbf{E} = 0 \quad (11)$$

The condition for propagation is thus seen to

reduce to the question of the existence of non-trivial fields, i.e., $\mathbf{E} \neq 0$. The necessary and sufficient condition for this is simply the vanishing of the determinant

$$\Delta(\mathbf{n}, \omega, \dots) = 0 \quad (12)$$

which expresses an algebraic relation between the wave normal direction \mathbf{n} (defined through $\mathbf{k} = k\mathbf{n}$, k a complex scalar) and the impressed frequency ω of the wave field. The other factors not specifically indicated are functions of the parameters of the medium itself. This algebraic relation is often called a "dispersion relation," since it is a relation for ω . It is only when this condition is fulfilled that a propagating plane wave field is possible. This condition also implies an equation for the refractive index of the medium. The same considerations that led to this relation also lead to a relation between the displacement field \mathbf{D} and the electric field \mathbf{E} , in the form

$$\mathbf{D} = \epsilon_0 \epsilon' \mathbf{E} \quad (13)$$

where ϵ' is a tensor whose structure for a particular choice of coordinate system (applied magnetic field co-directional with the x_3 -axis) is

$$\epsilon' = \begin{pmatrix} \epsilon_1 & j\epsilon_2 & 0 \\ j\epsilon_2 & \epsilon_1 & 0 \\ 0 & 0 & \epsilon_3 \end{pmatrix} \quad (14)$$

Here, the ϵ_j ($j = 1, 2, 3$) are complex valued and are functions of the parameters of the medium. If the medium is magneto-ionic (ions, electrons, and neutral molecules) and if collisions take place only between the charged species and the neutrals comprising the medium, the ϵ_j 's are rational functions of the medium parameters and the dispersion relation is also called the Appleton-Hartree equation, both of whom derived the relation independently.

The fact that the dispersion relation of such a medium depends upon the direction of the wave normal direction and ω makes the fourth-degree algebraic equation defining the refractive index complicated in practice. Theoretically, two indices are defined; the medium is doubly refracting in this case, analogous but certainly not identical with propagation in certain crystals, and this plasma crystal optics analogy can be advantageously exploited.* The propagating wave fields associated with each index are called, respectively, the *ordinary* mode, because it is least effected by the applied magnetic field, and the *extraordinary* mode. These two modes are linearly independent i.e., one of them cannot be expressed as a simple multiple of the remaining one. Both modes, in passing through the medium undergo attenuation and rotation of their respective planes of polarization; the so-called FARADAY EFFECT is present here as in some crystals. This effect is used in electron-density determinations, since the density

can be expressed in terms of the angle of rotation of these planes. The planes of polarization of each mode rotate opposite to each other.

Two important parameters (characteristic frequencies) are involved in discussing propagation: (1) the *plasma* frequencies of the constituents ω_j , defined by

$$\sqrt{N_j e^2 / \epsilon_0 m_j},$$

where subscript j stands for the type of species (electron, ion, etc.), N_j is the number density, m_j its mass, and e the electronic charge; and (2) the gyro or *Larmor* frequency defined by $\Omega_j = B_0 e / m_j$, where B_0 is the magnitude of the applied magnetic field and ω is the impressed wave frequency. The plasma frequency is essentially a resonant frequency and is more closely tied to the concept of collective motion or plasma oscillations. The Larmor frequency is a well-known concept and is substantially the frequency with which a charged particle "winds" itself about the externally applied magnetic field in the absence of collisions. If $\omega \neq \omega_j$ for all j , propagation is possible; if the medium consists only of electrons, the above condition is well known in ionospheric radio wave propagation.

Plasma Waves. For a more complicated description of the medium, other modes of propagation are possible. If the Lorentz force equation is extended so as to include pressure gradients (temperature effects are naturally included in the description by simple use of the gas laws), a hydrodynamic model of the gas results; and since this model is coupled to the electrodynamic equations through the Lorentz force term, we speak of magnetohydrodynamics or (to be more inclusive) magneto-fluid dynamics and its various synonyms (see MAGNETO-FLUID-MECHANICS). To see the effect of this coupling, we should recall that a completely incompressible perfect fluid with no magnetic field imposed cannot support any wave-like disturbances, as is obvious on physical grounds. This result is also clear on mathematical grounds in which the velocity potential satisfies Laplace's equation whose solutions are certainly not wave-like. If now the fluid is still incompressible but conducting and with a magnetic field imposed upon it, wave phenomena are now possible; the resulting waves are called magnetohydrodynamic and include both longitudinal and transverse types.

In the high-frequency limit, these waves yield the usual radio waves mentioned above; in the low-frequency limit, the resultant waves are no different in principle. However, because of the coupling of the more extensive hydrodynamic model with the electrodynamic field equations, the intermediate frequency region is naturally more complicated. The conditions for propagation (i.e., the dispersion relations) can no longer be readily discussed in general. The coupling, on the other hand, opens up a vast new area of investigation of wave phenomena, the existence of plasma waves, which was investigated initially by the astrophysicist Alfvén. In the case of an unbounded homogeneous medium, it is possible

* See, for instance, J. Brandstatter, "An Introduction to waves, Rays and Radiation in Plasma Media", New York, McGraw-Hill Book Co. (1963)

to treat, in a systematic way, the propagation of plane wave fields.

The conclusions drawn from the extensive hydrodynamical model in the low-frequency limit are also obtainable from a simpler point of view. For a magneto-ionic medium and $\omega \ll \omega_p$, Astrom and others deduced the properties of plane wave propagation by elementary considerations and arrived at the same result as Alfvén. The energy associated with the extraordinary mode proves to be propagated along the magnetic field lines. This manner of propagation sheds light on whistlers, a low-frequency natural phenomenon that exhibits similar properties. It is also worthwhile noting that since the dielectric tensor is in general non-symmetric, the condition of reciprocity (interchange of transmitter and receiver) does not hold as in the isotropic case in which ϵ reduces to a single scalar quantity. Finally, in the homogeneous unbounded case, the mean Poynting vector is *not* co-directional with the wave normal direction of either mode.

Nonhomogeneous Media. In nonhomogeneous, nonmagnetic media whose dielectric tensor varies in only one direction, the problem of calculating the wave fields can be reduced to the calculation of a Hertzian vector potential function, which satisfies an integral-equation relation. Here, one can employ variational techniques similar to those used by Schwinger, Levine, and others to find approximate solutions. Numerical techniques based on matrix theory have likewise been exploited. General reflection and transmission matrices are by-products of this problem. Other attempts to deal with this kind of problem, particularly for a doubly refracting medium separated from free space by a horizontal plane, are based on the ideas of Booker *et al.* A plane wave penetrating the medium from free space below splits into an ordinary and extraordinary wave. Because of the inhomogeneity of the medium, the penetrating waves which start at the plane interface generate along their path two ascending and two descending waves, the latter because of postulated internal reflections. The four waves so produced have different wave normal and Poynting vector directions. Snell's law is valid for both the ordinary and extraordinary waves. The inhomogeneous character of the medium causes the four waves, each of different elliptical polarization, to be coupled to each other. K. Suchy has shown that in the general nonhomogeneous medium, with no assumptions of stratification, Maxwell's equations reduce to a pair of coupled equations which has its simplest structure for a certain class of curvilinear coordinates. He has also treated the propagation of electromagnetic waves in absorbing, anisotropic, inhomogeneous, and unbounded media and studied in some detail the transition from the full vector wave equation with arbitrary wavelength λ to small values of λ but nonvanishing.

Ray-theoretic Approach. The study of propagation in anisotropic, inhomogeneous media becomes amenable to analysis by means of a

ray theory. Under the assumption that the wavelength λ is very much smaller than some characteristic dimension of the medium, it becomes possible to replace the more complicated vector field equations defined through the Maxwell-Lorentz system by a much simpler structure. By analogy with the homogeneous case, \mathbf{E} is taken to be proportional to

$$A(\mathbf{r}) \exp \left[j \left(\omega t - \omega \int \mathbf{M} \cdot d\mathbf{r} \right) \right]$$

where \mathbf{M} is in general a complex-valued vector. The \mathbf{H} field is expressed similarly. If this expression is used in the vector wave equation for \mathbf{E} we are led to a system with the following structure:

$$(\mathbf{L} - \epsilon)\mathbf{E} = 0 \quad (15)$$

where \mathbf{L} is a matrix whose elements contain the differential operators $\partial^2/\partial x_i \partial x_j$, and are complex. The condition for propagation is still the vanishing of the determinant $|\mathbf{L} - \epsilon|$, which is now a nonlinear partial differential equation of the second order for the three components of \mathbf{M} . The matrix \mathbf{L} is known as the "Eikonal matrix," and the above dispersion equation constitutes a generalization of the usual Eikonal equation or equation of geometrical optics. To understand the significance of the Eikonal equation, we consider a scalar wave equation whose spatial dependence f is of the form,

$$\nabla^2 f + \left(\frac{\omega M}{c} \right)^2 f = 0 \quad (16)$$

By taking $f(\mathbf{r}) = A(\mathbf{r}) \exp [- 2\pi j \psi(\mathbf{r})]$ and setting both the real and imaginary parts to zero, we obtain two nonlinear partial differential equations for the determination of both A and ψ . If both of these functions vary slowly within a distance of a wavelength, these equations simplify and one can find ψ without having to also find A . The resulting expression under these conditions for ψ is the Eikonal equation

$$4\pi^2 (\nabla \psi)^2 = \left(\frac{\omega M}{c} \right)^2 \quad (17)$$

whereas that for A takes the form

$$\nabla \cdot (A^2 \nabla \psi) = 0 \quad (18)$$

We can now easily relate ψ to \mathbf{M} ; indeed we put

$$\psi = \psi_1 - j\psi_2, \quad \mathbf{M} = \boldsymbol{\mu} - j\boldsymbol{\chi} \quad (19)$$

and in general both ψ and \mathbf{M} are functions also of frequency. The complex phase ψ is defined through

$$\psi = \int_{\mathbf{r}_0}^{\mathbf{r}} \nabla \psi \cdot d\mathbf{r} = \frac{\nu}{c} \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{M} \cdot d\mathbf{r} \quad (20)$$

where $d\mathbf{r}$ is a real vector and ν the frequency. The vector $\boldsymbol{\mu}$ is normal to the family of surfaces $\psi_1 = \text{constant}$, and $\boldsymbol{\chi}$ is normal to the family

$\psi_2 = \text{constant}$; these surfaces are in general distinct, and therefore we put

$$2\pi\psi_1 = \frac{\omega}{c} \int_{p_0}^p \mu \cdot d\mathbf{r} \quad (21)$$

$$2\pi\psi_2 = \frac{\omega}{c} \int_{p_0}^p \chi \cdot d\mathbf{r} \quad (22)$$

To find expressions for the magnitude of μ and χ we have

$$M^2 = (\mu - j\chi) = P - jQ \quad (23)$$

which is equivalent to

$$\mu^2 - \chi^2 = P; \quad 2\mu\chi \cos \theta = Q \quad (24)$$

where θ is the angle between μ and χ . These two relations yield

$$\begin{aligned} \mu^2 &= \frac{1}{2} \{ [P^2 + (Q/\cos \theta)^2]^{1/2} + P \} \\ \chi^2 &= \frac{1}{2} \{ [P^2 + (Q/\cos \theta)^2]^{1/2} - P \} \end{aligned} \quad (25)$$

the first expression defines a refractive index surface, and the second, an extinction index surface.

The equation defining A suggests a conservation law (equation of continuity) as applied to a fictitious fluid. By analogy we consider $A^2 \nabla \psi$ proportional to the current density of this fluid, and an application of the divergence theorem gives

$$\oint_V \nabla \cdot (A^2 \nabla \psi) dV = \oint_S A^2 \nabla \psi \cdot \mathbf{n} dS \quad (26)$$

where V is some volume containing the fluid and bounded by a surface S with outward drawn normal \mathbf{n} . If we consider this fluid to have (material) density proportional to A^2 and velocity proportional to $\text{grad } \psi$, then the above expression states that A^2 is conserved in time. If we assume that $|A|^2$ is a measure of the energy localized in the wave at each point, that $\text{grad } \psi$ is proportional to the energy flux at each point and is directed along the ray associated with the wave, then it appears as if the energy is a fluid which is conserved as it flows along the rays. Thus, the rays appear as trajectories of energy, a thin pencil of rays being analogous to a tube through which the energy flows.

In a homogeneous, anisotropic, but non-absorbing medium the concept of ray is defined as the trajectory whose direction is the same as that of the mean Poynting vector. This direction, as mentioned previously, is not the same as the wave normal direction. The task of defining the ray direction becomes more difficult for a non-homogeneous, anisotropic, absorbing, and dispersive medium. It can be shown that starting with the concept of a pulse in such a medium, one can derive extensions or generalizations of the usual Fermat's principle. This principle states that in a nonabsorbing, isotropic medium characterized by an index μ_r , a function of position only, the path of a so-called light ray makes the integral of μ_r between two fixed points

assume a stationary value. The pulse concept for the general case leads to the following stationary principles:

$$\begin{aligned} \delta \int \mu_r \cdot d\mathbf{r} &= 0 \\ \delta \int \chi \cdot d\mathbf{r} &= 0 \end{aligned} \quad (27)$$

$$\delta \int \frac{\partial}{\partial \omega} (\omega \mu) \cdot d\mathbf{r} = 0$$

where δ is the symbol of variation. In the non-absorbing case the first of Eqs. (27) gives the phase path, which is always normal to the surfaces of constant phase. Similarly, the trajectory of the amplitude surfaces is also normal to the surfaces of constant amplitude. For the absorbing case we retain these two definitions. This means that μ_r is always tangent to a path element of the phase trajectory and χ is tangent to a path element of the amplitude trajectory. These path elements are denoted, respectively, by $d\mathbf{r}_p$ and $d\mathbf{r}_a$. The pulse principle permits us to define, in addition to the above surfaces, the surfaces of constant group amplitude. These surfaces are distinguished from those of constant wave amplitude and constant wave phase. If $d\mathbf{r}_g$ denotes an element of the path along which the group amplitude propagates, then the stationary principle for this path takes the form

$$\delta \int \mu_r \cdot d\mathbf{r}_g = \delta \int \frac{\partial}{\partial \omega} (\omega \mu_r) \cdot d\mathbf{r}_g = 0 \quad (28)$$

and in general the direction of the vector $\partial/\partial \omega (\omega \mu_r)$ is not parallel to $d\mathbf{r}_g$. The vector $\partial/\partial \omega (\omega \mu_r)$ is the basis for the definition of the group velocity. By definition, the displacement of the center of the wave group in a time dt is

$$\mathbf{v}_g dt = d\mathbf{r}_g \quad (29)$$

where \mathbf{v}_g is the group velocity. If we define a vector $(\mathbf{v}_g)^{-1}$ (symbolic meaning only), such that

$$\mathbf{v}_g \cdot (\mathbf{v}_g)^{-1} = 1 \quad (30)$$

it is not difficult to show that

$$(\mathbf{v}_g)^{-1} = \frac{1}{c} \frac{\partial}{\partial \omega} (\omega \mu_r) \quad (31)$$

We take the path of energy propagation to be the path of group amplitude. This condition is certainly plausible when we consider that the energy of a field is usually proportional to the square of the amplitude of its oscillations.

In the case of an electromagnetic field, it is not so clear that the group path is also the energy path if this energy is measured by the mean Poynting vector, and the conditions under which this holds require careful analysis.

The question naturally arises as to how one actually calculates a ray path in the general case. The answer has been available for a long time. Hamilton's work in optics actually provides a

modus operandi for the practical calculation of the ray paths. Fermat's principle provides the starting point; we assume that the medium is characterized by an index function $m(\mathbf{r}, \alpha, \omega)$ called the ray refractive index. This function depends on position \mathbf{r} , direction α , and frequency ω , and is a homogeneous function of the first degree in α . The rays are defined as those curves such that

$$\delta \int_{r_0}^r m ds = 0 \quad (32)$$

where s is the parameter of arc length. The anisotropy is defined through the direction α which is normalized so that

$$\alpha \cdot \alpha = \frac{dr}{ds} \cdot \frac{dr}{ds} = 1 \quad (33)$$

that is, α defines the tangent to the ray. In the isotropic case this dependence is lacking and $m(\mathbf{r})$ is invariant with respect to proper rotations. Usually we do not know $m(\mathbf{r}, \alpha, \omega)$ but rather $\mu_r(\mathbf{r}, \mathbf{n}, \omega)$ where \mathbf{n} is the wave normal direction, as in the Appleton-Hartree equation. Thus it is desirable to attempt to express the ray path through the phase refractive index μ_r . This is accomplished by constructing a Hamiltonian defined by

$$H(\mathbf{r}, \sigma, \omega) = \frac{|\sigma|}{\mu_r(\mathbf{r}, \mathbf{n}, \omega)} = 1 \quad (34)$$

which is a homogeneous function of the first degree in σ , where $\sigma = \partial m / \partial \mathbf{r}$ is a vector in the direction of \mathbf{n} . It can be shown that the rays which satisfy Fermat's principle with m as index are identical with those satisfying

$$\delta \int_{r_0}^r \sigma \cdot d\mathbf{r} = 0 \quad (35)$$

whose variations are subject to the constraint $H(\mathbf{r}, \sigma, \omega) = 1$.^{*} The latter is Hamilton's principle and is fully equivalent to Fermat's principle. It follows from Hamilton's principle that the rays satisfy Hamilton's canonical equations

$$\frac{d\mathbf{r}}{dt} = \frac{\partial H}{\partial \sigma}, \quad \frac{d\sigma}{dt} = - \frac{\partial H}{\partial \mathbf{r}} \quad (36)$$

where t is a parameter along the path. Equations (36) are six simultaneous first-order nonlinear ordinary differential equations. Subject to initial conditions these equations have a unique solution. The advantage of defining the rays in terms of the first-order system as above are many, particularly from a computational point of view. The rays can be calculated for rather general media, and no assumptions of stratification are needed. The form of Hamilton's equations can easily be found for any coordinate system by tensor methods. Finally, we mention that a systematic study of ray propagation is possible without alluding to wave concepts although the ray-wave duality is implicit.

J. J. BRANDSTATTER

^{*} Brandstatter, *loc. cit.*

Cross-references: ELECTROMAGNETIC THEORY, FARADAY EFFECT, MAGNETO-FLUID-MECHANICS, PLASMAS, REFRACTION.

PROTON*

Protons in Atoms. The proton is the atomic nucleus of the element hydrogen, the second most abundant element on earth. Positively charged hydrogen atoms or "protons" were identified by J. J. Thomson in a series of experiments initiated in 1906. Although the structure of the hydrogen atom was not correctly understood at that time, several properties of the proton were determined. The electric charge on the proton was found to be equal but opposite in sign to that of an electron, and the measured value for the mass was much greater than that of the electron. The currently accepted proton mass is 1836 times the electron rest mass, or $1.672 \cdot 10^{-24}$ grams.

A correct estimate of the size of the proton and an understanding of the structure of the hydrogen atom resulted from two major developments in atomic physics: the Rutherford scattering experiment (1911) and the Bohr model of the atom (1913). Rutherford showed that the nucleus is vanishingly small compared to the size of an atom. The radius of a proton is the order of 10^{-13} cm as compared with atomic radii of 10^{-8} cm. Thus, the size of a hydrogen atom is determined by the radius of the electron orbits, but the mass is essentially that of the proton.

In the Bohr model of the hydrogen atom, the proton is a massive positive point charge about which the electron moves. By placing quantum mechanical conditions upon an otherwise classical planetary motion of the electron, Bohr explained the lines observed in optical spectra as transitions between discrete quantum mechanical energy states. Except for hyperfine splitting, which is a minute decomposition of spectrum lines into a group of closely spaced lines, the proton plays a passive role in the mechanics of the hydrogen atom. It simply provides the attractive central force field for the electron.

The proton is the lightest nucleus with atomic number one. Other singly charged nuclei are the deuteron and the triton which are nearly two and three times heavier than the proton, respectively, and are the nuclei of the hydrogen isotopes deuterium (stable) and tritium (radioactive). The difference in the nuclear masses of the isotopes accounts for a part of the hyperfine structure called the "isotope shift."

In 1924, difficulties in explaining certain hyperfine structures prompted Pauli to suggest that a nucleus possesses an intrinsic angular momentum or "spin" and an associated magnetic moment. The proton spin quantum number (I) is $1/2$, and the angular momentum is given by $[I(I+1)h^2/(2\pi)^2]^{1/2}$ where h is Planck's constant.

^{*} Supported in part by The U.S. Air Force Office of Scientific Research.

The intrinsic magnetic moment is 2.793 in units of nuclear magnetons (0.50504×10^{-23} erg/gauss) which is about a factor of 660 less than the magnetic moment of the electron.

Two types of hydrogen molecule result from the two possible couplings of the proton spins. At room temperature, hydrogen gas is made up of 75 per cent orthohydrogen (proton spins parallel) and 25 per cent parahydrogen (proton spins antiparallel). Several gross properties, such as specific heat, strongly depend on the ortho or para character of the gas.

Protons in Nuclei. Protons and neutrons are regarded as "nucleons" or fundamental constituents of nuclei in most theories of nuclear structure and reactions. The nuclear forces operating between them are much stronger than the electrostatic forces which govern atomic and molecular systems but operate over very short ranges, the order of several times 10^{-13} cm. Of particular significance in the structure of nuclei is the apparent charge independence of the forces. That is, the nuclear force between two nucleons may be considered separately from the electrostatic forces due to electric charges the nucleons may carry. In addition to mass and charge, other properties such as spin and parity play important roles in determining the mechanics of nuclei.

The mass of a nucleus is the sum of the masses of the nucleons contained, plus a correction due to the total binding energy of the nucleons. This correction is an application of the Einstein mass-energy equivalence ($E = mc^2$). The atomic number or positive charge of a nucleus is given by the number of protons.

A detailed description of the motion of a proton in a nucleus is complicated by the many-body, quantum mechanical nature of the problem, but several simplified theories or models have been very successful in predicting many of the properties of nuclei. One of the best known nuclear structure models is the shell model in which a nucleon is assumed to move in a central force field. This field represents the average interaction of the proton (or neutron) with all other nucleons. An essential additional assumption is a coupling of the orbital angular momentum of the independent nucleon with its spin.

The success of a theory in which a nucleon moves among its close-packed neighbors as if they were not present is due in part to quantum mechanical restrictions, in particular to the Pauli exclusion principle.

The optical model for the scattering of protons by nuclei also rests on the assumption that the interaction with many nucleons may be represented by an average potential well. An imaginary potential term is included which accounts for reactions other than elastic scattering.

Much of nuclear physics may be understood with a picture in which the proton exists as an independent particle in the nucleus. Refinements of an independent particle model to include collective effects and deformation involve a consideration of the residual interaction between

nucleons and the details of the individual nucleon interaction with the average potential well.

Certain aspects of the very strong nuclear forces are understood. These forces involve π mesons in somewhat the same way that electrostatic forces responsible for atomic structure involve photons. Yukawa introduced the π meson as the field quantum for nuclear forces in 1935, and the interaction potential derived from this early theory is commonly used in nuclear physics.

Structure of the Proton. A major objective of physics is to identify elementary particles and determine their properties. The proton is important in these investigations in several connections. It is itself an elementary particle. It is used as a projectile in the production and study of other elementary particles. It has been used as a target in the study of nucleon structure.

One view is to consider all fundamental particles as states of excitation of a limited number of particles. The resulting simplification correlates a large body of experimental data.

The internal structure of the proton is still being studied. In one of the experiments, the proton is a target which is probed by very short-wave length electrons. Electrons produced by a high-energy linear accelerator are scattered from protons to study the electric charge distribution in a proton.

Questions concerning the structure of the proton or the neutron are difficult due to the ambiguity of experimental results. One cause of uncertainty is that the probe is not defined in sufficient detail. It appears desirable to use a strong interaction probe, that is, another nucleon or meson, but the structure of the probe may be no better understood than the structure of the target.

A mantle of mesons has been found about the proton. A current question is whether or not there is an internal or intrinsic core of a proton.

R. H. DAVIS

References

- Shortley, George, and Williams, Dudley, "Elements of Physics," Englewood Cliffs, N. J., Prentice-Hall, 1961.
- Born, Max, "Atomic Physics," Glasgow, Blackie & Son, Ltd., 1946.
- Weidner, Richard T., and Sel's, Robert L., "Elementary Modern Physics," Boston, Allyn and Bacon, 1960.
- Richtmyer, F. K., Kennard, E. H., and Lauritsen, T., "Introduction to Modern Physics," New York, McGraw-Hill, 1955.

Cross-references: ACCELERATORS, PARTICLE; ATOMIC AND MOLECULAR BEAMS; ATOMIC PHYSICS; ELECTRON; ELEMENTARY PARTICLES; ISOTOPES; NEUTRON; NUCLEAR STRUCTURE; QUANTUM THEORY; RADIOACTIVITY; RELATIVITY.

PULSE GENERATION

Pulse waveforms are distinguished by abrupt, often almost discontinuous features. Such features may serve as time markers: pulses can carry information in the time domain (e.g., in computers). Also, such waveforms can have a very low duty cycle (= fractional time the pulse is present); pulses can supply momentary large bursts of power to equipment whose average rating is relatively low (e.g., high-power microwave pulse transmitters for radar).

The fundamental element in pulse generation is the switch. Opening or closing a switch generates a step-function pulse (e.g., in key telegraphy); one or more such steps, together with suitable wave-shaping operations, may yield the final desired waveform.

For low-level pulses, overdriven amplifiers (transistors or vacuum tubes) are commonly used as switching elements. The simplest example is that of a *clipping amplifier*; when driven by a large sinusoid, this produces an approximately square-wave output whose abrupt level changes may serve as sources of pulse waveforms. The transition *rissetime*, beyond its dependence on the driving signal, is limited by amplifier properties such as internal capacitances and charge storage (in transistors). Low-power transistors may have rissetimes of the order of nanoseconds; high-power units are typically an order of magnitude slower.

Two amplifiers can be cross-connected to drive each other alternately over the switching range. Depending on the coupling circuits, the pair can be bistable, monostable, or astable (Fig. 1). The *univibrator* rests in its quiescent state until driven into the other by the trigger pulse; it relapses into the rest state after a time determined by the R-C coupling, approximately given by $RC \ln(1 + V_{cc}/V_{bb})$. Thus the trigger produces an output of standardized amplitude and width. The *multivibrator*, with two R-C couplings, free-runs and thus serves as pulse source (clock). Timing (or frequency) stabilities of order 1 per cent are typical.

The *blocking oscillator* is a single amplifier which drives itself over the switching range by means of transformer feedback. A monostable arrangement is shown in Fig. 2. The transistor is

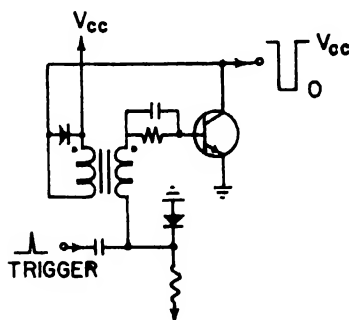


FIG. 2. Monostable blocking oscillator.

brought momentarily into conduction by the trigger pulse; thereafter it drives itself into saturation. The width of the output pulse is determined mostly by the transformer properties, above all by the magnetic saturation of the core. The regenerative self-drive permits the blocking oscillator to supply relatively large currents to the load.

Steps and rectangles can be shaped into other forms by many techniques.¹ Three classes (linear, nonlinear, and delay-line shaping) are typified by the examples in Fig. 3. The delay-line circuit produces a narrow rectangle from a step; the trailing edge is formed by the reflection of the leading edge in the unmatched delay line. The pulse falltime is thus as short as the rissetime, within the limitations of the delay line response.

Whenever possible, switching and shaping are carried out directly at the output power level. In that case, the energy in the delivered pulse may come from that stored in the pulse-forming network, the switch dissipation being negligible (current and voltage not present simultaneously);

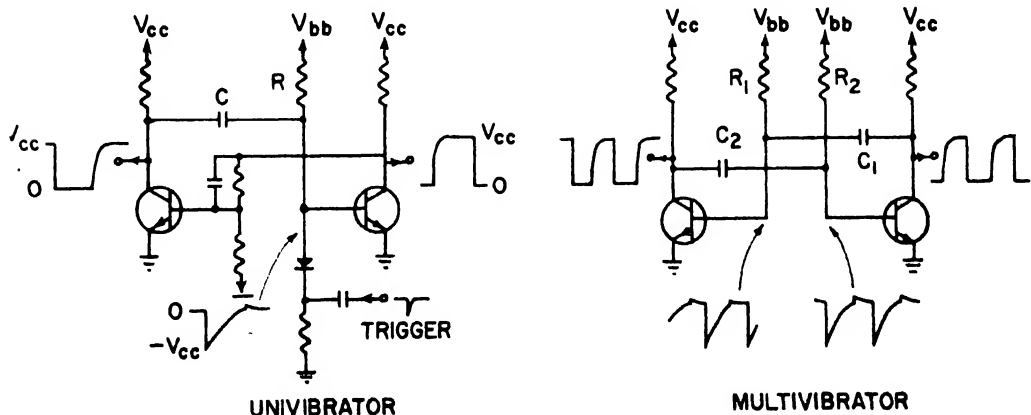


FIG. 1. Univibrator (monostable) and multivibrator (astable). If both cross-couplings are resistive, a bistable binary results; this requires a trigger on either side to change state.

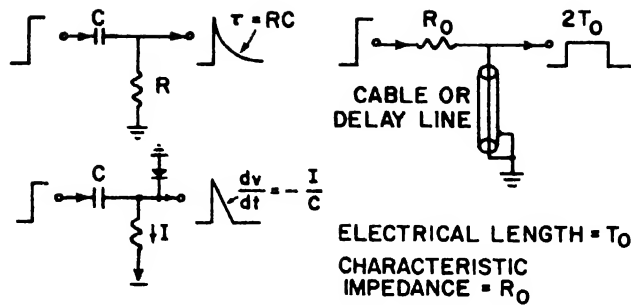


FIG. 3. Examples of shaping circuits to produce short pulses from a step.

the power supply recharges the network slowly between pulses. For some shapes, the forming process is inefficient and must be carried out at low level; a linear output amplifier is then required. This, as well as the supply, must momentarily handle large power. Vacuum-tube amplifiers are common in this type of service (voltages up to about 50 kV, currents to several tens of amperes).

Some other types of switching element are listed below:

(a) *Mechanical Switches* (usually reed relays with mercury-wetted contacts to eliminate bounce) may generate steps with sub-nanosecond risetime. The electromechanical drive limits the frequency to a few hundred cycles per second.

(b) *Unijunction Transistors*² (double-base diodes) switch regeneratively when the emitter becomes forward biased; the injected carriers increase the conductivity of the base material. Switching times of the order of 1 μ sec are typical.

(c) *Thyratrons* (hot-cathode gas- or vapor-filled tubes). A discharge is initiated by a trigger signal applied to the grid. To extinguish the discharge, the anode must be reverse-biased; the deionization time (of order 100 μ sec) limits the minimum pulse spacing. Extremely high power levels may be handled: up to about 50 kV, 5 kA. The risetime is as short as a few nanoseconds in hydrogen tubes. Figure 4 shows a pulse modulator for a microwave klystron. The

power stored in the pulse-forming network is delivered to the load via a pulse transformer. Recharge takes place through a "ringing choke," avoiding the power loss associated with a charging resistor.

(d) *Ignitrons, Triggered Spark Gaps*. These are high-power gaseous discharge switches. Spark gaps are simple and produce extremely short risetimes, but electrode erosion may be a problem.

(e) *Controlled Rectifiers*³ (pnpn devices) have operating characteristics in many ways similar to those of thyristors. In most types, the load current must be interrupted to restore the non-conducting state; some low power devices (controlled switches) can be turned off by their control electrode. Risetimes are typically of order 1 μ sec.

(f) *Avalanche Transistors*⁴. Collector-to-emitter breakdown may occur regeneratively with sub-nanosecond risetime in certain transistors. This may be initiated by overvoluting the collector or by suddenly changing the base potential. Avalanche transistors can generate fast, large pulses (e.g., 30 volts) at high repetition rates (e.g., 5 Mc/sec).

(g) *Tunnel Diodes*⁵. Heavily doped diodes have extremely thin depletion layers through which quantum-mechanical tunneling of majority charge carriers can occur over a narrow voltage range (about 0.1 volt). At higher voltage, after a sharp drop of this tunnel current, ordinary injection

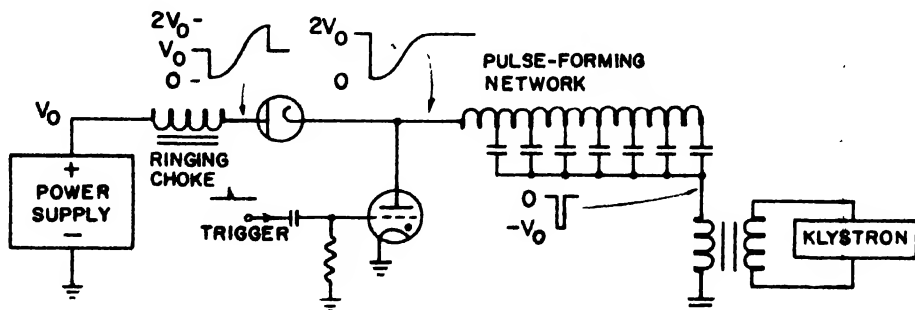


FIG. 4. High-power pulser using hydrogen thyatron.

current sets in (about 0.5 volt). The diode may switch from tunneling to injection voltage at sub-ns speed. Choice of loadline results in bistable, monostable, or astable operation; transitions may be provoked by suitable current signals. The voltage level of operation is very low (about 0.5 volts).

(h) *Snap Diodes*⁸. Stored charge resulting from a period of forward conduction causes a diode to conduct freely when reverse bias is first applied. In special step-recovery diodes, this reverse current ceases abruptly (with sub-nanosecond falltime), and the resulting opening of the switch can generate fast pulses at high repetition rates (100 Mc/sec) and relatively high level (tens of volts).

R. M. LITTAUER

References

1. Strauss, L., "Wave Generation and Shaping," New York, McGraw-Hill Book Company, 1960.
2. "Transistor Manual," Sixth edition, Ch.13, General Electric Company, Syracuse, N. Y., 1962.
3. "Silicon Controlled Rectifier Manual," Second edition, General Electric Company, Auburn, N. Y., 1961.
4. Henebry, W. M., "Avalanche Transistor Circuits," *Rev. Sci. Instr.*, **32**, 1198 (1961).
5. "RCA Tunnel Diode Manual," Radio Corporation of America, Somerville, N. J., 1963.
6. Moll, J. L., *et al.*, "P-N Junction Charge Diodes," *Proc. IRE*, **50**, 43 (1962); Giorgis, J., "Understanding Snap Diodes," *Application Note*, **90.17** (1963), General Electric Company, Syracuse, N. Y.

Cross-references: DIODE (SEMICONDUCTOR), ELECTRICITY, ELECTRON TUBES (RECEIVING TYPE), FEEDBACK, RADAR, TRANSISTOR, TUNNELING.

PYROMETRY, OPTICAL

Optical pyrometry is that branch of thermometry in which the temperature of a solid or liquid is determined by measuring the radiation emitted in a relatively narrow spectral region. The International Practical Temperature Scale (IPTS) above the melting point of gold (1336.15°K) is defined in terms of such radiation, and an optical pyrometer is usually used to realize the IPTS above this temperature.

Figure 1 is a schematic diagram of a common type of optical pyrometer. The objective lens (B) images that part of the source (A) for which the temperature is to be determined in the plane of the filament of the pyrometer lamp (E). The microscope (G, H, I) magnifies the source image and the small pyrometer lamp filament by a factor of about fifteen. The red filter (F) and the spectral response of the eye limit the spectral band pass of the instrument to wavelengths from about 6200 to 7100Å with the effective value about 6500Å. The absorption filter (D) is inserted at source brightness temperatures above about 1300°C so that the pyrometer lamp

need not be operated at higher temperatures where its stability is poor.

The optical pyrometer shown in Fig. 1 is operated by adjusting the current in the pyrometer lamp filament until the brightness of the filament equals the brightness of the image of the source.

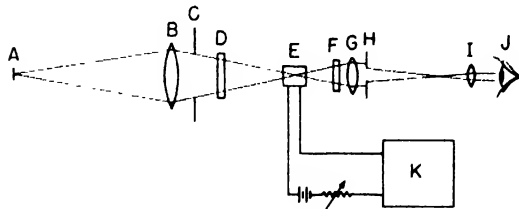


FIG. 1. Schematic diagram of a disappearing filament optical pyrometer. A. Source; B. Objective lens; C. Objective aperture; D. Absorption filter (used for temp. above 1300°C); E. Pyrometer lamp; F. Red filter; G. Microscope objective lens; H. Microscope aperture stop; I. Microscope ocular; J. Eye; K. Current measuring instrument.

From this current and a previous calibration of the pyrometer, the brightness temperature of that part of the source or object sighted upon can be determined. The brightness temperature of an object is defined as the temperature of a blackbody which emits the same spectral radiance as the object. Spectral radiance is defined as the quotient of the radiant power emitted at a particular wavelength and in a particular direction by the product of the wavelength interval, the solid angle and the emitting area projected perpendicular to the direction of sighting. If the object is a blackbody, the brightness temperature of the blackbody is its temperature. If the object is not a blackbody, the temperature of the object can be obtained from its brightness temperature and the equation

$$\epsilon_{\lambda} = \frac{e^{C_2/\lambda T} - 1}{e^{C_2/\lambda T_B} - 1} \quad (1)$$

where ϵ_{λ} is the spectral emissivity (also called spectral emittance), λ is the wavelength, in centimeters, at which the pyrometer is effectively operating (usually about 0.65×10^{-4} cm), C_2 is the second radiation constant (1.438 centimeter degrees on the IPTS), T_B is the brightness temperature and T is the temperature, both in degrees Kelvin. The spectral emissivity of a surface is the fraction of blackbody spectral radiance emitted by the surface when it has the same temperature as the blackbody. Spectral emissivities for a wavelength of 0.65×10^{-4} cm have been determined for a large number of materials. When $C_2/\lambda T$ is greater than about 5, Eq. (1) can be replaced, with negligible error, by the simpler equation

$$\frac{1}{T} = \frac{1}{T_B} + \frac{\lambda}{C_2} \ln \epsilon_{\lambda} \quad (2)$$

where \ln is the symbol for the natural logarithm.

The calibration of an optical pyrometer from basic principles is called a primary calibration and is usually performed in national standard laboratories such as the National Bureau of Standards. The first step of a primary calibration of an optical pyrometer is making a brightness match while sighting on a blackbody surrounded by freezing gold. The temperature of freezing gold is defined to be 1336.15°K on the International Practical Temperature Scale. Higher brightness temperature points are obtained by using the defining equation of the IPTS,

$$R = \frac{e^{c_2/\lambda T_n} - 1}{e^{c_2/\lambda T_{Au}} - 1} \quad (3)$$

Experimentally, the brightness temperature of a stable source is adjusted until the source, as seen through a rotating sector disk with transmittance R , has the same brightness as a gold blackbody. Independent measurement of R and λ then permit, from Eq. (3), a calculation of T_n . The pyrometer lamp current required for making a brightness match while sighting on the source without the sector disk completes the calibration at this higher temperature. This process is repeated with a sufficient number of sector disks and temperatures so that a smooth curve can be drawn relating current in the pyrometer lamp to the brightness temperature of the source.

With a well-designed optical pyrometer of the type in Fig. 1, it is possible to realize the IPTS in a primary calibration with an estimated uncertainty of ± 0.4 deg at 1336°K, ± 2.0 deg at 2300°K and ± 10.0 deg at 4300°K. The National Bureau of Standards also calibrates commercial pyrometers of the type in Fig. 1 by comparison to a primary calibrated pyrometer. The uncertainty of this comparison calibration is estimated to be ± 3 deg, ± 6 deg, and ± 40 deg respectively at the temperatures given above. These figures apply to brightness or blackbody temperatures. The accuracy with which an optical pyrometer can determine the temperature of a non-blackbody depends not only on the pyrometer calibration in terms of brightness temperature but also on the uncertainty of the spectral emissivity required. This often produces an error greater than the calibration uncertainties given above.

Recently (1960), photoelectric pyrometers have been developed which replace the eye with a photomultiplier tube and use an interference filter or monochromator to limit the spectral band pass. With the increased sensitivity and smaller band pass in these instruments, the International Practical Temperature Scale above 1336.15°K can be realized with about one-third the uncertainty possible with the visual instrument in Fig. 1. The present limitation in realizing the IPTS with a photoelectric pyrometer appears to be the instability of the pyrometer lamps. To improve the accuracy further, lamps with greater stability will have to be developed or a different type optical pyrometer designed, possibly one using only the gold-point blackbody as a reference source. The gold-point blackbody is at least a factor of ten more stable than pyrometer lamps.

In addition to optical pyrometers, total radiation pyrometers and two color pyrometers are sometimes used for measuring temperatures. Usually, however, these instruments are less accurate than optical pyrometers and are primarily intended for controlling temperature rather than for measuring it.

H. J. KOSTKOWSKI

References

- Kostkowski, H. J., and Burns, G. W., "Thermocouple and Radiation Thermometry Above 900°K, Measurement of Thermal Radiation Properties of Solids (1963)," Office of Scientific and Technical Information, National Aeronautics and Space Administration, U.S. Government Printing Office, Washington, D. C.
- Kostkowski, H. J., and Lee, R. D., "Theory and Methods of Optical Pyrometry," *Natl. Bur. Std. Monograph*, 4, (1962).
- Forsythe, W. E., "Optical Pyrometry," in *Temperature, its Measurement and Control in Science and Industry*, p. 115, New York, Reinhold Publishing Corp., 1941.
- "Temperature, Its Measurement and Control in Science and Industry," Vol. III, Parts 1 and 2, New York, Reinhold Publishing Corp., 1962.

Cross-references: TEMPERATURE AND THERMOMETRY; RADIATION, THERMAL.

Q

QUANTUM ELECTRODYNAMICS

Broadly speaking, quantum electrodynamics is that branch of physics which studies the phenomena of interaction of radiation with matter subject to the laws of quantum mechanics. In recent years, it has been studied extensively not only to explain the results of experiments with photons and elementary particles and nuclei, but also to obtain a deeper understanding of the basic laws of physics. However, progress has been very slow, undoubtedly on account of the complexity of the problem and the lack of knowledge of the structure of elementary particles and the interacting forces at close ranges which account for their stability and the creation or annihilation of material particles and radiation. On the other hand, there is a greater understanding of the basic laws governing pure radiation and of its interaction with electrons and positrons. The theory of the electromagnetic fields and the electron-positron and their interaction subject to quantum laws are well understood, and in a proper sense, quantum electrodynamics dealing with Maxwellian and Dirac fields, free or in interaction with each other, is as complete as one could expect from our present knowledge of the electromagnetic field and the electron-positron pair, based on experimental evidence.

Quantization of Free Maxwellian Fields. In quantizing the electromagnetic field, one starts with the laws of classical electrodynamics based on Maxwell's equations. These may be written in the form

$$F_{\mu\nu,\rho} + F_{\nu\rho,\mu} + F_{\rho\mu,\nu} = 0 \quad (1a)$$

$$F_{\mu\nu,\nu} = 0 \quad (1b)$$

where $F_{\mu\nu}$ are the well-known components of the field, E and H . They can be derived also from a variational principle by taking a Lagrangian

$$L = -\frac{1}{2}F_{\mu\nu}F_{\mu\nu} \quad (2)$$

Usually, one works with the potentials A_μ^* , defined by

$$F_{\mu\nu} = A_{\nu,\mu} - A_{\mu,\nu} \quad (3)$$

* The Greek letters run from 1 to 4 for vectors or operators and the Latin indices from 1 to 3. In the coordinates, the Greek subscript takes the values 0, 1, 2, 3, and Latin letters, 1, 2, 3. Also $k = (\mathbf{k}, k_0) = (k_1, k_2, k_3, k_0)$, $x = (\mathbf{x}, x_0) = (x_1, x_2, x_3, x_0)$, the same notation holds for p, q ; $k_0 = \omega$ (omega), $x_0 = it$ and the units are $\hbar = c = 1$.

which satisfy the equation of propagation (motion)

$$\square A_\mu(x) - A_{\nu,\mu\nu}(x) = 0 \quad (4)$$

From Eq. (3) A_μ is determined up to a gradient of a scalar function $\Lambda(x)$, i.e.,

$$A_\mu^1 = A_\mu + \Lambda_{,\mu} \quad (5)$$

known as gauge transformation. The fields $F_{\mu\nu}$ and Eq. (4) remain invariant under such transformations.

The equation of motion [Eq. (4)] is simplified by introducing a subsidiary (Lorentz) condition on the potentials, namely

$$A_{\mu,\mu} = 0 \quad (6)$$

reducing Eq. (4) to

$$\square A_\mu = 0 \quad (7)$$

which together with Eq. (6) is equivalent to Maxwell's equations. However, condition (7) still does not determine uniquely A_μ , but now the gauge transformation restricts the class of functions Λ to solutions of the equation

$$\square \Lambda = 0 \quad (8)$$

On account of the photon mass $m = 0$, the fields $F_{\mu\nu}$ and not the potentials A_μ have a direct physical meaning. This, however, is not the case if $m \neq 0$.

The canonical (momenta) variables $\pi_\mu(x)$ corresponding to the Lagrangian Eq. (3) are

$$\pi_\mu = iF_{4\mu}(x) \quad (9)$$

and thus $\pi_4 = 0$. As a consequence, Eq. (9) cannot be solved by $A_{\mu,4}$. To avoid this difficulty, the Lagrangian Eq. (3) is replaced by the following

$$L = -\frac{1}{2}F_{\mu\nu}F_{\mu\nu} - \frac{1}{2}A_{\mu,\mu}A_{\nu,\nu} = -\frac{1}{2}A_{\mu,\nu}A_{\nu,\mu} \quad (10)$$

from which the canonical momenta are given by

$$\pi_l(x) = iF_{4l}(x) \quad (11a)$$

$$\pi_4(x) = iA_{\nu,\nu}(x) \quad (11b)$$

and hence instead of Eq. (8) we obtain the weaker condition

$$\square A_{\nu,\nu}(x) = 0 \quad (12)$$

which follows from Eq. (4). Furthermore, one does not take the general solution of the equation,

but only those which satisfy the initial conditions, say at $t = 0$,

$$A_{r,r}(x) = 0, A_{r,0r}(x) = 0, \text{ for all } x \quad (13)$$

then from Eq. (12) it follows that $A_{r,r}(x)$ vanishes for all t . This is in agreement with the usual formalization of the theory of electromagnetism.

To quantize the field, we take the Lagrangian Eq. (10) and apply the commutation rules to the canonical variables $A_\mu(x)$, $\pi_\mu(x)$ for $x_0 \rightarrow x'_0$, namely

$$\begin{aligned} [A_\mu(x), A_\nu(x')] &= 0, [\pi_i(x), \pi_k(x')] \\ &= A_{i,0}(x) - A_{k,0}(x') = 0 \\ [A_\mu(x), \pi_k(x')] &= [A_\mu(x), A_{k,0}(x')] \\ &= i\delta_{\mu k}\delta(\mathbf{x} - \mathbf{x}') \quad (14) \\ [A_4(x), \pi_4(x')] &= [A_4(x), iA_{r,0}(x')] \\ &= i\delta_{44}\delta(\mathbf{x} - \mathbf{x}') \end{aligned}$$

Now we expand $A_\mu(x)$ in a Fourier series in momentum space k

$$A_\mu(x) = \sum_k [e^{ikx} A(\mathbf{k}) + e^{-ikx} A(\mathbf{k})] \quad (15)$$

From Eq. (7) it is necessary that $k^2 = \mathbf{k}^2 = k_0^2$, so $k_0 = \pm \omega$, $\omega = \sqrt{\mathbf{k}^2}$. On the other hand, since $A_i(x)$ is hermitian and $A_4(x)$ antihermitian, it follows that $A_i(k)$, $iA_4(k)$ and $A_i(\mathbf{k})$, $iA_4(\mathbf{k})$ are hermitian conjugates of each other. Furthermore, for every \mathbf{k} , four possible (independent) polarizations e_μ^λ , $\lambda = 1, 2, 3, 4$ (orthogonal) can be assigned to $A_\mu(x)$, so Eq. (15) may be written in the form

$$A_\mu(x) = \frac{1}{\sqrt{2V}} \sum_k \sum_\lambda \frac{e_\mu^\lambda}{\omega} [e^{ikx} a^\lambda(\mathbf{k}) + e^{-ikx} a^{*\lambda}(\mathbf{k})],$$

$$(e_\mu^\lambda e_\mu^{\lambda'} - \delta_{\lambda\lambda'}) (k_0 = \omega) \quad (16)$$

In general e_μ^λ may be functions of k . We then choose e_μ^λ in such a way that

$$e_4^m = 0, e_m^1 k_m = e_m^2 k_m = 0, e_m^3 = k_m \omega^{-1},$$

$$e_m^4 = 0, e_i^4 = 1, (m = 1, 2, 3) \quad (17)$$

With this choice of the polarization vector e_μ^λ , its components for $\lambda = 1, 2$ represent transverse polarization, $\lambda = 3$, longitudinal (along \mathbf{k}) polarization and $\lambda = 4$ scalar polarization. Since condition (6) has not been imposed on $A_\mu(x)$, we also have $\sum_\lambda e_\mu^\lambda e_\nu^{\lambda'} = \delta_{\mu\nu}$ and from the choice $e_i^4 = 1$, $a^\lambda(\mathbf{k})$, $a^{*\lambda}(\mathbf{k})$ satisfy the same conditions as $A_\mu(\mathbf{k})$ and $A_\mu^*(\mathbf{k})$.

With $A_\mu(x)$ given by Eq. (17), the commutation relations of Eq. (14) now become

$$[a^\lambda(\mathbf{k}), a^{\lambda'}(\mathbf{k}')] = [a^{*\lambda}(\mathbf{k}), a^{*\lambda'}(\mathbf{k}')] = 0$$

$$[a^\lambda(\mathbf{k}), a^{*\lambda'}(\mathbf{k}')] = \delta_{\lambda\lambda'} \delta_{\mathbf{k}\mathbf{k}'} \quad (18)$$

by using the properties of e_μ^λ and the definition of the delta function.

From the Lagrangian function Eq. (10) one finds, after partial integration in \mathbf{x} , the Hamiltonian operator H to be

$$H = \frac{1}{2} \sum_{\mathbf{k}, \lambda} \{a^\lambda(\mathbf{k}), a^{*\lambda}(\mathbf{k})\} \omega \quad (19)$$

where $\{a, a^*\} = aa^* + a^*a$, symbolizes anticommutation. Introducing hermitian operators $q_\mathbf{k}^\lambda$, $p_\mathbf{k}^\lambda$ defined by

$$q^m = \frac{1}{\sqrt{2\omega}} (a^m + a^{*m}), p^m = i\sqrt{\frac{\omega}{2}} (a^m - a^{*m})$$

$$(m = 1, 2, 3)$$

$$q^4 = \frac{1}{\sqrt{2\omega}} (a^4 - a^{*4}), p^4 = i\sqrt{\frac{\omega}{2}} (a^4 + a^{*4}),$$

$$[p^\lambda, q^{\lambda'}] = i\delta_{\lambda\lambda'} \quad (20)$$

H takes the form

$$H = \frac{1}{2} \sum_{\mathbf{k}} \left\{ \sum_{\lambda=1}^3 [(p^\lambda)^2 + (q^\lambda)^2] + [(p^4)^2 + (q^4)^2] \right\} \quad (21)$$

which resembles the Hamiltonian operator for a sum (infinite in our case) of independent harmonic oscillators. Therefore, the energy operator H and its eigenvalues are

$$H = \sum_{\mathbf{k}} \left\{ \sum_{m=1}^3 N^m(\mathbf{k}) + N^4(\mathbf{k}) \right\} \omega,$$

$$E = \sum_{\mathbf{k}} \omega \left\{ \sum_{m=1}^3 n^m(\mathbf{k}) + n^4(\mathbf{k}) \right\},$$

$$(N^m = a^{*m} a^m, N^4 = a^4 a^{*4}) \quad (22)$$

In Eq. (22) the zero-point energy of the oscillators (infinite) has been neglected. It represents the polarization of the vacuum, which for free fields does not play an important role. However, this is not true if sources or interaction with particles are considered.

A similar calculation leads to the following expression for the moment of momentum

$$P_m = \frac{1}{2} \sum_{\mathbf{k}} \{a^\lambda(\mathbf{k}), a^{*\lambda}(\mathbf{k})\} k_m$$

$$= \sum_{\mathbf{k}} k_m \left\{ \sum_{\lambda=1}^3 N^\lambda(\mathbf{k}) + N^4(\mathbf{k}) \right\} \quad (23)$$

where the zero-point contribution has been neglected. Equations (12) and (23) show the corpuscular aspect of the electromagnetic field. However, they include negative terms, i.e., both E and P_m may take negative values, in contradiction to the classical theory. Furthermore, H includes, besides transverse photons ($\lambda = 1, 2$), longitudinal ($\lambda = 3$) and scalar photons ($\lambda = 4$), in contradiction to our concept of light. In a correct theory, the last two kinds of photons should not appear in Equations (22) to (23).

Before describing the method of eliminating longitudinal and scalar photons, we shall briefly discuss the case when the commutation rules are

valid at different times. In this case, the time part in the expressions do not cancel, so instead of Eq. (14), we have the following result:

$$[A_\mu(x), A_\nu(x')] = \frac{1}{V} \sum_{\mathbf{k}, \lambda} \frac{e_{\mu\lambda} e_{\nu\lambda}}{2\omega} [e^{ik(x-x')} - e^{-ik(x-x')}] \\ = -i\delta_{\mu\nu} D(x' - x) \quad (24)$$

where the singular function $D(x)$ is defined by

$$D(x) = i(2\pi)^{-3} \int e^{ikx} \delta(k^2) \epsilon(k) dk \\ \delta(k^2) = \delta(k^2 - k_0^2) = \frac{1}{2|k|} [\delta(|\mathbf{k}| - k_0) + \\ + \delta(|\mathbf{k}| + k_0)], \quad \epsilon(k) = \frac{k_0}{|k_0|} \quad (25)$$

The function $D(x)$ satisfies the following properties:

$$D(x) = 0 \text{ for } x^0 > 0, \quad D(x)_{x_0} = -\delta(\mathbf{x}) \text{ for } x_0 = 0, \\ \square D(x) = 0 \quad (26)$$

This form of commutation rule is relativistic invariant, making the commutator of two fields a c -number. If interactions with particles or external fields are included, the result is a q -number.

Let $|0\rangle$ be the state vector of null particles. Then the expectation value of the anticommutator $\{A_\mu(x), A_\nu(x')\}$ is a matrix given by

$$\langle 0 | \{A_\mu(x), A_\nu(x')\} | 0 \rangle = (\delta_{\mu\nu} - \delta_{\mu 4} \delta_{\nu 4}) D^1(x' - x) \\ D^1(x) = (2\pi)^{-3} \int e^{ikx} \delta(k^2) dk \quad (27)$$

With the aid of $D(x)$ and $D^1(x)$ one can construct the corresponding functions for the retarded, advanced and mixed potentials which play important roles in quantum electrodynamics. All these functions satisfy the non-homogeneous equation $\square D_\mu = -\delta_\mu(x)$. In fact, when the mass $m \neq 0$, $D(x)$ is modified by replacing $\delta(k^2)$ by $\delta(k^2 + m^2)$ in Eq. (25).

In our derivation of Eq. (21) and (22), we have not used Lorentz condition (6), since it is incompatible with the commutation relations of Eq. (14). The potentials $A_\mu(x)$ or the $a^\lambda(\mathbf{k})$, $a^{*\lambda}(\mathbf{k})$ are operators acting on state functions $|\psi\rangle$; the Lorentz condition must be modified so that it is compatible with the commutation rules. This is done by requiring that

$$\langle A_\nu, a(x) | \psi(x) \rangle = 0 \quad (6')$$

$$\langle \psi | A_{\nu,r}(x) | \psi \rangle = 0 \text{ (expectation value of } A_{\nu,r}) \quad (6'')$$

However, since the state functions $|\psi\rangle$ cannot be normalized, one introduces a "metric operator" η and define the norm of a state vector to be: Norm $|\psi\rangle = \langle \psi | \eta | \psi \rangle$. As the norm should be a real quantity, η must be hermitian, $\eta^* = \eta$ and one assumes $\eta^2 = \eta \eta^* = 1$. Similarly the expectation value of any operator H is expressed in the form

$$H^{**} = \langle \psi | \eta H | \psi \rangle \quad (28)$$

From the new definition of normalizing the state functions and from Eq. (28) one deduces

$$\langle \psi | \eta A_k(x) | \psi \rangle = \langle \psi | A_k^*(x) \eta^* | \psi \rangle = \\ = \langle \psi | A_k(x) \eta | \psi \rangle \\ \langle \psi | \eta A_4(x) | \psi \rangle = \dots = \langle \psi | A_4 \eta | \psi \rangle \quad (28a)$$

We find for the commutators and anticommutators in ordinary and momentum space

$$[A_m(x), \eta] = 0, \quad [a^\lambda(\mathbf{k}), \eta] = 0, \quad \lambda \neq 4 \\ \{A_4(x), \eta\} = 0, \quad \{a^4(\mathbf{k}), \eta\} = 0 \quad (28b)$$

It follows from these relations that $\langle n^\lambda(\mathbf{k}) | \eta | n'^\lambda(\mathbf{k}) \rangle = \delta_{nn'}$, $\lambda \neq 4$ and

$$\langle n^4 | \eta | n'^4 \rangle = \delta_{nn'} (-1)^{n^4}$$

Applying the subsidiary condition Eq. (6') by using the operators $a^\lambda(\mathbf{k})$, we have

$$[a^3(k) + ia^4(k)] | \psi \rangle = 0 \quad (28c)$$

or in k -space (28) is written as

$$A^+_{\nu,r}(x) | \psi \rangle = 0 \quad (29)$$

where $A^+(x)$ represents the positive frequency part of $A(x)$, i.e., the term with e^{ikx} . Denoting $A^-(x)$, the negative frequency part of $A(x)$, we must also have $\langle \psi | A_{\nu,r}^-(x) | \psi \rangle = 0$. Equation (6'') for the expectation value of $A_\mu(x)$ now becomes

$$\langle \psi | \eta A_{\nu,r}(x) | \psi \rangle = 0 \quad (30)$$

The new subsidiary condition, now allows many independent solutions of Eq. (28). Any state vector $|\psi\rangle$ can be represented in the form

$$|\psi\rangle = |\psi_T\rangle \Pi \phi_k \quad (31)$$

Inserting this expression in Eq. (28), we obtain a fundamental solution by choosing

$$|\phi^0\rangle = |0, 0\rangle, \quad |\phi^1\rangle = |1, 0\rangle + i|0, 1\rangle, \dots, \\ |\phi^n\rangle = \sum_{r=0}^n i^r \sqrt{\binom{n}{r}} |n-r, r\rangle, \dots \quad (32)$$

All the vectors $|\phi^n\rangle$ are orthogonal to each other and their norms are not negative. In fact all, the norms of $|\phi^n\rangle$ vanish, except for that with $n=0$. Thus we have

$$\langle \phi^n | \eta | \phi^{n'} \rangle = 0, \quad n \neq n', \quad \langle \phi^n | \eta | \phi^n \rangle = \delta_{n0} \quad (33)$$

To satisfy Eq. (28) we normalize $|\phi^0\rangle$ to one and so we can express the vectors

$$|\phi_k\rangle = |\phi_k^0\rangle + \sum_{n \neq 0} c^n(\mathbf{k}) |\phi^n(\mathbf{k})\rangle \quad (34)$$

where the $c^n(\mathbf{k})$ are arbitrary. In terms of a^λ , Eq. (28) yields the following relation

$$a^3 |\phi^n\rangle = \sqrt{n} |\phi^{n-1}\rangle, \quad a^4 |\phi^n\rangle = i \sqrt{n} |\phi^{n-1}\rangle,$$

and the expectation values of $a^3(\mathbf{k})$, $a^4(\mathbf{k})$ reduce to the simple expressions

$$\langle \phi_k \eta a^3(\mathbf{k}) | \phi_k \rangle = C^1(\mathbf{k}), \quad \langle \phi_k \eta a^4(\mathbf{k}) | \phi_k \rangle = i C^1(\mathbf{k}) \quad (35)$$

On the other hand, the expectation value of A_μ is, from Eq. (5), given by

$$\langle \psi | \eta A_\mu(x) | \psi \rangle = \Lambda_{,\mu}(x) \quad (36)$$

Then on account of Eq. (36), Λ may be expanded in the form

$$\Lambda(x) = \frac{i}{\sqrt{V}} \sum_k \frac{1}{\sqrt{2}\omega^3} [c^{*1}(\mathbf{k})e^{-ikx} - c^1(\mathbf{k})e^{ikx}] \quad (37)$$

where the coefficients c^1 and c^{*1} are chosen so that Eq. (37) satisfies Eq. (8). Therefore, the expansion of a state vector according to Eq. (33) eliminates all the difficulties resulting from the Lorentz condition (6) and also the presence of scalar and longitudinal photons in the energy of the field, since the terms with coefficients $c^a(\mathbf{k})$ do not contribute any eigenstate to the Hamiltonian operator. The expectation value of the energy is now given only in terms of transverse photons, since only the state vectors $|\phi^0\rangle$ contribute to the Hamiltonian operator. The expectation value of H has now the form

$$\langle \psi | \eta H | \psi \rangle = \sum (n_k^1 + n_k^2) \omega \quad (38)$$

The vacuum state is made up of a mixture of longitudinal and scalar photons only, i.e., the vacuum is a state devoid of particles, (photons) which we have denoted by the state vector $|0\rangle$.

Quantization of the Free Dirac Field. The quantizations of the free Dirac field can be carried out in a similar manner as for the Maxwell free field, by making appropriate modifications for the inclusion of a finite mass and charge. The equation of motion of free electrons is the Dirac equation

$$(\gamma_\mu \partial_\mu + m)\psi_\mu(x) = 0 \quad (39)$$

where γ_μ are Dirac matrices, γ_4 being expressed by Pauli matrices and $\gamma_4 = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}$. With this choice of γ_μ , and the adjoint $\bar{\psi}(x) = \psi^*(x)\gamma_4$, where $\psi^*(x)$ is the hermitian adjoint of ψ , $i\bar{\psi}(x)\gamma_\mu\psi(x)$ is invariant.

We now express a solution of Eq. (39) in a plane wave form

$$\psi_\mu(x) = u_\mu(\mathbf{q})e^{i(\mathbf{q}\mathbf{x} - q_0x_0)} \quad (40)$$

Substituting this in Eq. (39) one obtains four independent solutions $u_{\mu r}(\mathbf{q})$, two solutions, $r = 1, r = 2$, for positive q_0 and two for $r = 3, r = 4$ for $-q_0$, where $q_0 = \pm E = \pm \sqrt{\mathbf{q}^2 + m^2}$. These solutions are normalized in such a way that

$$\sum_{\mu=1}^4 u_{\mu r}(\mathbf{q})u_{\mu s}(\mathbf{q}) = \delta_{rs} \quad (41)$$

$$\sum_{r=1}^4 u_{\mu}^{*r}u_{rr} = \delta_{\mu\nu}, \sum_{r=1}^2 \bar{u}_{\mu}^r u_{rr} = -(2E)^{-1}(i\gamma q^+ - m)\nu_\mu, \quad (42)$$

$$\sum_{r=3}^4 \bar{u}_{\mu}^{-r} u_{rr} = (2E)^{-1}(i\gamma q^- - m)\nu_\mu$$

where $\bar{u}(\mathbf{q}) = u^*(\mathbf{q})\gamma_4$, and $q^+ = (\mathbf{q}, E)$, $q^- = (\mathbf{q}, -E)$. Equations (41) and (42) are the orthogonality and completeness relations.

The Lagrangian, the Hamiltonian and the current density of the fields are of the form

$$L = -\bar{\psi}(x)\left(\gamma \frac{\partial}{\partial x} + m\right)\psi(x) \quad (43)$$

$$H = -i\pi_\psi(x)\gamma_4(\gamma_\lambda \partial_\lambda + m)\gamma(x) \quad (44)$$

$$j = ie\bar{\psi}(x)\gamma_\mu\psi(x) \quad (45)$$

In the theory of quantization of the Dirac field, the functions $\psi(x)$, $\bar{\psi}(x)$ are considered as operators acting on state functions unlike that of the ordinary theory, where ψ is a state function. This procedure is known as second quantization of the electrons and $\bar{\psi}(x)\psi(x)$ does not have a probability interpretation as in the ordinary Dirac theory.

Developing the field functions $\psi(x)$ in plane waves, as was carried out for Maxwellian fields, in q -space we write

$$\psi_\mu(x) = \frac{1}{\sqrt{V}} \sum_q \left\{ e^{i(\mathbf{q}\mathbf{x} - Et_0)} \sum_{r=1}^2 u_{\mu r}(\mathbf{q})a^r(\mathbf{q}) + e^{i(\mathbf{q}\mathbf{x} + Et_0)} \sum_{r=3}^4 u_{\mu r}(\mathbf{q})a^r(\mathbf{q}) \right\} \quad (46)$$

and substituting it in Eq. (44) and (45), we obtain for the Hamiltonian operator and charge Q

$$H = \sum_q E \left\{ \sum_{r=1}^2 a^{*r}(\mathbf{q})a^r(\mathbf{q}) - \sum_{r=3}^4 a^{*r}(\mathbf{q})a^r(\mathbf{q}) \right\} \quad (47)$$

$$Q = e \sum_q \sum_{r=1}^4 a^{*r}(\mathbf{q})a^r(\mathbf{q}) \quad (48)$$

To find the commutation relations for the $a^r(\mathbf{q})$, $a^{*r}(\mathbf{q})$, one makes use of the property that the commutator of H with a field operator F is equal to $-iF$. From Eqs. (46) and (47), we find the following anticommutation rules

$$\{a^{*r}(\mathbf{q}), a^s(\mathbf{q})\} = \delta_{rs}\delta_{\mathbf{q}\mathbf{q}'}, \{a^{*r}(\mathbf{q}), a^{*s}(\mathbf{q}')\} = 0 \\ \{a^r(\mathbf{q}), a^s(\mathbf{q}')\} = 0 \quad (49)$$

Now the total energy can be written in terms of the particle numbers, N^+ , N^- ,

$$H = \sum_q E \left\{ \sum_{r=1}^2 (N^{+r}(\mathbf{q}) - N^{-r}(\mathbf{q}) - 2) \right\} \\ N^{+r} = a^{*r}(\mathbf{q})a^r(\mathbf{q}), N^{-r} = b^{*r}(\mathbf{q})b^r(\mathbf{q}), \\ b^1(\mathbf{q}) = a^{*4}(-\mathbf{q}), b^2(\mathbf{q}) = a^{*3}(-\mathbf{q}) \quad (50)$$

It is not difficult to see that $N^+(1 - N^+) = N^-(1 - N^-) = 0$. The last term of Eq. (50) represent an infinite energy, the so-called zero-point energy of the field. Therefore, the eigenvalues of N^+ and N^- are 0 and 1, and from the anticommutation relations of Eq. (49), for any given state there is only one electron in agreement with Pauli's principle.

In a similar manner, one finds the following expression for the total charge

$$Q = e \sum_{\mathbf{q}, r} \{N^+{}^r(\mathbf{q}) - N^-{}^r(\mathbf{q})\} \quad (51)$$

where the zero-point part has been neglected. This expression for the charge is not positive definite. The interpretation given to N^+ and N^- is that for the positive part of the energy, and negative e , N^+ denotes the number of electrons, and for the negative part of the energy, and $+e$, N^- represents the number of positrons.

The zero-point term in Eq. (51) when formulated in x -space must yield a value for the expectation of the current density $j_\mu(x)$ for the vacuum.

This can be done by taking $j_\mu(x) = \frac{ie}{2} [\bar{\psi}(x) \gamma_\mu \psi(x)]$ instead of Eq. (45). Using Eq. (46), one arrives directly at Eq. (51) and also obtains $\langle 0 | j_\mu(x) | 0 \rangle = 0$, where $|0\rangle$ is the state vector of the vacuum.

The ordinary components of the momentum are obtained from the expansion of Eq. (46), as follows:

$$P_k = i \int d^3x \bar{\psi}(x) \gamma_k \psi(x)_{,k} = \sum_{\mathbf{q}, r} \hbar \mathbf{q} \cdot (N^+{}^r(\mathbf{q}) - N^-{}^r(\mathbf{q})) \quad (52)$$

From the definition of b^{\pm} Eq. (50), the last term in Eq. (46) is the creation operator for the positrons. The angular momenta can also be calculated from the expression of the energy-momentum tensor $T_{\mu\nu}(x)$. Its space components are made up of terms which are independent of the polarization state of the particles (ordinary angular momenta), and terms along the direction of motion for any electron state. If z is the direction of motion, then for an electron state $|q\rangle$, and position state $|q'\rangle$, the latter terms are

$$J_{12}{}^1 |q\rangle = \frac{1}{2} (-1)^{r-1} |q\rangle, \quad J_{12}{}^1 |q'\rangle = \frac{1}{2} (-1)^{r-1} |q'\rangle \quad (53)$$

In the expression for the charge [Eq. (51)], N^+ and N^- correspond to the electron and positron particle if e is taken negative and have a reverse role if e is assigned a positive value.

Therefore the theory is invariant with respect to the transformation $N^+ \leftrightarrow N^-$, and at the same time $+e \leftrightarrow -e$. This invariant property can be formulated also for the field operators $\psi(x)$ in x -space. Remembering that $\bar{\psi}(x)$ is the annihilation operator for the electrons, but the creation operator for the positron, and the reverse for $\bar{\psi}(x)$ or $\psi(x)$, the transformation $N^+ \leftrightarrow N^-$ corresponds to an interchange of $\bar{\psi}(x)$ and $\psi(x)$. This symmetry is not preserved for charge transformations, so the transformation with respect to charge reversal (conjugation) is expressed in a more complicated form.

$$\psi_A(x) \rightarrow \psi_B^+(x) = C_{AB} \bar{\psi}_B(x), \quad C_{AB} = -C_{BA},$$

$$(C^{-1})_{AB} = (C_{BA})^*, \quad (C^{-1} \gamma_\mu C)_{AB} = -(\gamma_\mu)_{AB} \quad (54)$$

These properties of C make it possible for $\bar{\psi}'(x)$ to satisfy the same equation as $\psi(x)$, leave $j_\mu(x)$ invariant, and $\psi'(x)$ can be developed in a series similar to Eq. (46). However, the Lagrangian [Eq. (43)] is no longer invariant and symmetric with respect to charge conjugation. To preserve these properties one introduces a new Lagrangian

$$L = \frac{1}{4} \left[\bar{\psi}(x) \left(\gamma \frac{\partial}{\partial x} + m \right) \psi(x) \right] - \frac{1}{4} \left[\frac{\partial \bar{\psi}(x)}{\partial x} \gamma + m(x) \right] \psi(x) \quad (55)$$

from which one obtains the correct equations of motion and the energy momentum tensor $T_{\mu\nu}$. By choosing the matrix $C = \gamma_2 \gamma_4$, then C is a unitary matrix, with the property, $C^{-1} \gamma_\mu C = \gamma_\mu$, $\mu = 2, 4$ and $+\gamma_\mu$ for $\mu = 1, 3$, and the vacuum expectation value of the current density, $\langle 0 | j_\mu(x) | 0 \rangle = 0$, is preserved. Furthermore, one can calculate in a similar way as for the electromagnetic case the commutators and anticommutators of $\bar{\psi}(x)$ and $\psi(x')$, in x -space using the development of Eq. (46). The result of summing over the index r leads, with the aid of Eq. (42), to the following expression for the anticommutator

$$\{\bar{\psi}_A(x), \psi_B(x')\} = -S_{BA}(x' - x) \quad (56)$$

$$S_{AB} = -i(2\pi)^{-3} \int dq e^{iqx} (i\gamma q - m)_{AB} \delta(q^2 + m^2) \epsilon(q) \quad (57)$$

where the singular function S plays the same role as $D(x)$ for Maxwellian fields, and moreover, it is a solution of Dirac's equation with initial conditions $\psi(x) = u(x)$ for $x_0 = T$. From S , we can construct the corresponding retarded and advanced functions S_R , S_A , and the mixed function S_F , satisfying Dirac's Eq. (1) with a non-homogeneous term $2i\delta(x)$. On the other hand, the vacuum expectation value of the commutator, $\bar{\psi}(x)$, $\psi(x)$ is expressed in terms of the singular function

$$S^1(x) = (2\pi)^{-3} \int dp e^{ipx} (i\gamma p - m) \delta(m^2 - p^2),$$

which does not vanish for space-like points, but vanishes for $x_0 = x'_0$ as required by the theory. The function $S^1(x)$ plays the same role as $D^1(x)$ in Maxwell theory.

Free Maxwellian and Dirac Fields in Interaction. One of the most important problems in quantum electrodynamics is that of interaction of electrons with radiation fields. One of the great successes of quantum electrodynamics is to be found in predicting and explaining a large number of phenomena related to behavior of electrons and photons in interaction, such as the Lamb shift, magnetic moment of the electron, and collision processes. These achievements are possible on account of the weak interaction between electrons and photons. However, the theory is beset by a number of serious difficulties, mainly due to the effect of the vacuum on the mass and charge of the electrons.

The equations of motion of an electron interacting with a Maxwellian field can be derived from a Lagrangian L

$$L = L_e + L_M + L_i \quad (58)$$

where L_e and L_M are the Lagrangians of the free Dirac and Maxwell field, and L_i is the interaction Lagrangian given by the expression

$$L_i = \frac{ie}{2} A_\mu(x) [\bar{\psi}(x), \gamma_\mu \psi(x)] - A_\mu(x) j_\mu(x) \quad (59)$$

From the Lagrangian function, one derives the equation of motion in the usual way

$$\left(\gamma \frac{\partial}{\partial x} + m \right) \psi(x) = ie \gamma A(x) \psi(x) \quad (60)$$

$$\square A_\mu(x) = \frac{ie}{2} [\bar{\psi}(x), \gamma_\mu \psi(x)] - j_\mu(x) \quad (61)$$

Since L_i does not contain time derivatives of the field operators, the quantization is carried out as previously. The commutation relations, the operators $A(x)$, $\psi(x)$, are unchanged as for free fields, but we have in addition

$$[A_\mu(x), \psi(x)] = [A_{\mu, x_0}(x), \psi(x')] = 0 \text{ for } x_0 = x'_0$$

$$[A_\mu(x), \bar{\psi}(x')] = [A_{\mu, x_0}(x), \bar{\psi}(x')] = 0 \quad (62)$$

However, one cannot extend the commutation relations to different times in a simple way as for the free fields, since Eqs. (60) and (61) do not have simple solutions and, as we have remarked, they no longer lead to c -numbers as for free fields.

The integration of the equations of motion (coupled equations) can be done by the method of Picard (successive approximation), providing the coupling parameter is small. In general, with the aid of the functions S_R and S_A , one can transform them to coupled integral equations, where S_R and S_A are the Green functions of the problem. The usual method is to apply perturbation theory by expanding $\psi(x)$, $A_\mu(x)$ in series of the coupling constant (fine structure constant), with the leading terms taken as solutions of the free field equations. This procedure is satisfactory as long as we do not take account of the effect of the vacuum, which introduces changes in the mass and the charge of the electron which are expressed by divergent integrals. By a process known as *renormalization*, the infinities entering in the expressions of variations in mass and charge, δm and δe , can be made to disappear, so that the total mass and charge are the physical measured quantities. In this way, meaningful results can be assigned to the elements of the matrix representing physical quantities.

The source of these infinities arises from the fluctuations of the vacuum states of the electromagnetic field and the electron-positron pairs, even though their expectation values vanish. Thus, the interaction of an electron with the vacuum states produces a change in its energy through the variation of mass (renormalization

effect) and also affects the electron-positron state, which in turn changes the state of polarization of the vacuum, manifested in an apparent change of charge of the electron (charge renormalization). Even though one is able to extract meaningful results from the renormalization procedure, it is found to be not unitary, and unless convergence is established in a unique way, negative norms may creep into the energy operator which would lead to meaningless negative probabilities. Another difficulty enters into the picture when strong interactions come into play (high energies) with the creation of mesons, since perturbation procedures fail. Even other mathematical methods (asymptotic procedures) have not yielded many useful results, on account of the mathematical complexity of the problem. Likewise for very small intervals, the geometry may no longer be Minkowskian or even Riemannian, and the present concept of space and time would have to be modified in order to make the theory free of the divergencies appearing in the integrals representing the operators and also in the renormalization process. In spite of these grave difficulties in the theory, that part of quantum electrodynamics dealing with photon and electrons, free and in interaction with each other, will not be affected very much by new developments.

NICHOLAS CHAKO

References

- Achieser, A. I., and Beresterki, W. B., "Quantenelektrodynamik," Frankfurt Main, 1962.
 Bogoliubov, N. N., and Shirkov, D. V., "Introduction to the Theory of Quantized Fields," New York, Interscience Publishers, New York, 1959.
 Henley, E. M., and Thirring, W., "Elementary Field Theory," New York, McGraw-Hill Book Co., 1962.
 Schweber, S. S., "An Introduction to Relativistic Quantum Field Theory," Elmsford, N.Y., Harper & Row, 1961.
 Kastler, D., "Introduction à l'Électrodynamique Quantique," Dunod, Paris, 1961 (modern mathematical approach)
 Kallen, G. "Handbuch der Physik," Vol. VI, Berlin, Springer, 1958.

Cross-references: ELECTROMAGNETIC THEORY, MATRIX MECHANICS, QUANTUM THEORY.

QUANTUM THEORY

Experiments on thermal radiation, with their gross disagreement with classical theories, gave birth to quantum theory. The equilibrium distribution of electromagnetic radiation (that is, emission and absorption of radiation at constant temperature) in a hollow cavity could not be explained on the basis of classical electrodynamics (Maxwell's equations plus the laws of motion of particles). Thermal radiation is a certain function of the temperature (T) of the emitting body. When dispersed by a prism, thermal radiation forms a continuous spectrum. It was found that

the energy distribution of the radiation had a regular dependence on its wavelength. Furthermore the energy E_ν as a function of the temperature of the material did not depend upon the structure of the cavity or its shape. On these bases it was shown that the energy E_ν ought to have a dependence upon frequency ν , at temperature T , in the form

$$E_\nu = \nu^3 F\left(\frac{cT}{\nu}\right)$$

All attempts to find the correct form of F on the basis of classical theory failed. The classical theory led to the now well-known "ultraviolet difficulty," since the contribution of high frequencies caused the energy to assume infinite value. The difficulty was removed by a hypothesis of Planck, according to which the energy of a monochromatic wave with frequency ν can only assume those values which are integral multiples of energy $h\nu$; that is, $E_n = nh\nu$, where n is an integer referring to the number of "photons." Thus the energy of a single PHOTON of frequency ν is

$$E = h\nu \quad (1)$$

The finiteness of Planck's constant h and its resulting implications laid the foundations of quantum theory. Quantum theory, like the special theory of relativity, was discovered through the experiments on electromagnetic phenomena and their theoretical interpretations.

The fundamental equation of quantum mechanics [Eq. (1)] implies, on the one hand, that energy of radiation stays concentrated in limited regions of space in amounts of $h\nu$ and, therefore, behaves like the energy of particles; on the other hand, it establishes a definite relationship between the frequency ν and the energy E of an electromagnetic wave. This dual behavior of light corresponds, in one way, to experimental situations of the interference properties of radiation, for the description of which one uses the wave theory of light; in another way, it corresponds to the properties of exchange of energy and momentum between radiation and matter, which require for their explanation the particle picture of light. Thus the dual behavior of light has necessitated the quantum description (quantization) of the electromagnetic field. A unified point of view was formulated quantitatively by de Broglie, according to which all forms of energy and momentum related to matter will manifest a dual behavior of belonging to a wave or particle description of the physical system, depending on the type of experiment performed.

The most interesting example of a quantum mechanical object is the photon itself. By using the relativistic and quantum mechanical definition of the photon energy, we can obtain a quantitative formulation of the above ideas. The relativistic form of the total energy of a particle with rest mass m and momentum p is

$$E = c\sqrt{(p^2 + m^2c^2)} \quad (2)$$

We set $m = 0$ and obtain the relativistic definition of the energy of a photon:

$$E = cp \quad (3)$$

Hence the first unification of relativity and quantum theory originated from the combination of Eqs. (1) and (3) in the form

$$cp = h\nu \quad (4)$$

By using $\nu\lambda = c$ for the plane electromagnetic wave, we obtain the fundamental statement of quantum mechanics,

$$\lambda p = h \quad (5)$$

valid for all particles with or without mass, where

$$\lambda = \frac{h}{p}, \quad h = \frac{h}{2\pi}$$

These assumptions of quantum theory have laid the foundations of new physical and philosophical concepts for the process of measurement in physics and the definition of physical reality.

We need to develop a dynamical theory to describe the wave character of material particles. We shall base our approach on the idea that the concept of the photon must play a fundamental role in building a quantum theory of matter. To this end, a preliminary understanding of free photons in quantum theory will provide a first orientation, and it will set a clear path for further generalizations of the subject matter.

In the case of particles with mass, one has the possibility of comparing their kinetic energies with their rest masses. If the kinetic energy is small compared to rest energy then we can formulate a nonrelativistic theory. However, with the photon there exists no possibility for the formulation of a nonrelativistic theory. The theory of a free photon will have to be a relativistic one; it is a relativistic particle. There are important advantages in entering quantum mechanics via the photon:

(a) The energy of a photon is a quantum mechanical quantity, $E = h\nu$.

(b) It has provided a natural basis to postulate the wave-particle relation, $\lambda = h/p$.

(c) The wave aspects of the photon are completely described by charge-free Maxwell equations. Therefore, it is natural to try to reconcile Planck's hypothesis with the wave theory of light.

A reinterpretation of Maxwell's equations in conjunction with the quantum relations $E = h\nu$ and $\lambda = h/p$ will, in a natural way, lead us to a wave equation for the photon.

We begin with the free-field Maxwell equations

$$\frac{1}{c} \frac{\partial \mathcal{E}}{\partial t} - \nabla \times \mathcal{H} = 0, \quad \frac{1}{c} \frac{\partial \mathcal{H}}{\partial t} + \nabla \times \mathcal{E} = 0 \quad (6)$$

$$\nabla \cdot \mathcal{E} = 0, \quad \nabla \cdot \mathcal{H} = 0 \quad (7)$$

Now consider the complex three-dimensional vector

$$\mathbf{X} = \mathcal{E} + i\mathcal{H} \quad (8)$$

It has the following interesting properties.

(a) For a plane electromagnetic wave, whose electric and magnetic vectors are perpendicular and equal (in magnitude), the Lorentz-invariant square of \mathbf{X} vanishes; i.e.,

$$(\mathcal{E} - i\mathcal{H})^2 = \mathcal{E}^2 - \mathcal{H}^2 - 2i\mathcal{E} \cdot \mathcal{H} = 0 \quad (9)$$

or

$$|\mathcal{E}| = |\mathcal{H}|, \quad \mathcal{E} \cdot \mathcal{H} = 0$$

The latter two are the two Lorentz-invariant properties of a plane electromagnetic wave.

(b) The energy density of the field can be expressed in terms of the rotation-invariant square of \mathbf{X} as

$$cP_4 = \frac{1}{8\pi} |\mathbf{X}|^2 = \frac{1}{8\pi} \langle \mathbf{X} | \mathbf{X} \rangle = \frac{\mathcal{E}^2 + \mathcal{H}^2}{8\pi} \quad (10)$$

where $|\mathbf{X}\rangle$ is the column vector

$$|\mathbf{X}\rangle = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

and $\langle \mathbf{X} | = [X_1^*, X_2^*, X_3^*]$ is Hermitian conjugate row vector of $|\mathbf{X}\rangle$.

(c) The momentum density of the field has the form

$$P_i = \frac{1}{8\pi c} \langle \mathbf{X} | K_i | \mathbf{X} \rangle \quad (11)$$

where the Hermitian 3×3 matrices K_i were defined by

$$K_j = \begin{bmatrix} 0 & -i\delta_{j3} & i\delta_{j2} \\ i\delta_{j3} & 0 & -i\delta_{j1} \\ -i\delta_{j2} & i\delta_{j1} & 0 \end{bmatrix} \quad (12)$$

where δ_{ij} is the usual Kronecker tensor with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

The two expressions of Eqs. (10) and (11) for energy and momentum densities, respectively, can be combined into a single equation,

$$cP_\mu = \frac{1}{8\pi} \langle \mathbf{X} | K_\mu | \mathbf{X} \rangle \quad (13)$$

where K_4 , corresponding to $\mu = 4$, is a 3×3 unit matrix. In order to illustrate the meaning of Eq. (13) we shall, as an example, work out the third component of Eq. (13). It is given by

$$\begin{aligned} cP_3 &= \frac{1}{8\pi} [X_1^*, X_2^*, X_3^*] \begin{bmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \\ &= \frac{i}{8\pi} (X_1 X_2^* - X_1^* X_2) = \frac{1}{4\pi} (\mathcal{E}_1 \mathcal{H}_2 \\ &\quad - \mathcal{E}_2 \mathcal{H}_1) = \frac{1}{4\pi} (\mathcal{E} \times \mathcal{H})_3 \end{aligned}$$

which verifies the statement in (c) above.

(d) The momentum and energy densities of the field are conserved; i.e.,

$$\frac{\partial cP_\mu}{\partial x_\mu} = \frac{\partial P_4}{\partial t} + c \nabla \cdot \mathbf{P} = 0 \quad (14)$$

provided the complex vector $|\mathbf{X}\rangle$ satisfies the equation

$$K_\mu \frac{\partial}{\partial x_\mu} |\mathbf{X}\rangle = 0 \quad (15)$$

Equation (15) can also be written as

$$i \frac{\partial}{\partial t} |\mathbf{X}\rangle = -ic \mathbf{K} \cdot \nabla |\mathbf{X}\rangle \quad (16)$$

where, it is easy to see,

$$-i \mathbf{K} \cdot \nabla |\mathbf{X}\rangle = \nabla \times \mathbf{X} \quad (17)$$

Hence Eq. (16) is equivalent to Maxwell's equations [Eq. (6)]. Therefore, the conservation of energy and momentum density of the field is a consequence of Maxwell's equations or, conversely, if Maxwell's equations are satisfied, the vector P_μ is conserved.

We write Maxwell's equations in the form

$$i\hbar \frac{\partial}{\partial t} |\mathbf{X}\rangle = H |\mathbf{X}\rangle \quad (18)$$

where

$$\nabla \cdot \mathbf{X} = 0 \quad (19)$$

$$H = -ic\hbar \mathbf{K} \cdot \nabla = c \mathbf{K} \cdot \mathbf{p} \quad (20)$$

has the dimensions of energy. The momentum operator \mathbf{p} is defined as

$$\mathbf{p} = -i\hbar \nabla$$

The complex vector $|\mathbf{X}\rangle$ has the physical dimensions of the square root of energy density. It will be more convenient to work with a complex vector $|\eta\rangle$, having the dimensions of the square root of (volume)⁻¹, defined by

$$|\mathbf{X}\rangle = \sqrt{(8\pi E)} |\eta\rangle \quad (21)$$

and satisfying the condition of normalization,

$$\int \langle \eta | \eta \rangle d^3x = 1 \quad (22)$$

where E is the energy of the photon. With these premises, the wave equation (or Schrödinger's equation) of the photon becomes

$$i\hbar \frac{\partial}{\partial t} |\eta\rangle = H |\eta\rangle \quad (23)$$

$$\nabla \cdot \eta = 0 \quad (24)$$

The formalism contained in Eqs. (23) and (24), with the definitions of H by Eq. (20), will be shown to be consistent and compatible with the observed facts $E = \hbar\nu$ and $\lambda = \hbar/p$ of the photon.

The Fourier series expansion of the wave function can be given as

$$|\eta\rangle = \frac{1}{\sqrt{V}} \sum_k e^{i\mathbf{k} \cdot \mathbf{r}} |\phi_k\rangle \quad (25)$$

where summation refers to all three components of the \mathbf{k} vector and is over all the k , defined according to

$$k_1 = \frac{2\pi}{L} n_1, \quad k_2 = \frac{2\pi}{L} n_2, \quad k_3 = \frac{2\pi}{L} n_3 \quad (26)$$

The complex vectors $|\phi_k(t)\rangle$ are functions of the wave number and of time t . By substituting the function $|\eta\rangle$ given by Eq. (25), in Eqs. (23) and (24), we obtain

$$i\hbar \frac{\partial}{\partial t} |\phi_k\rangle = H |\phi_k\rangle \quad (27)$$

$$\mathbf{k} \cdot \boldsymbol{\phi}_k = 0 \quad (28)$$

where n_1, n_2, n_3 are integral numbers specifying wave number vector \mathbf{k} in a plane wave $\exp(i\mathbf{k} \cdot \mathbf{r})$, and

$$H = \hbar c \mathbf{k} \cdot \mathbf{K}$$

so the operator p acts on $|\phi_k\rangle$ according to

$$p|\phi_k\rangle = \hbar \mathbf{k} |\phi_k\rangle \quad (29)$$

Thus the vector $\hbar \mathbf{k}$ can be interpreted as the eigenvalue of the operator p corresponding to the eigenvector $|\phi_k\rangle$.

In order to find the eigenvalues of the operator H , we may look for the stationary-state solutions of Eq. (27). We assume that the vector $|\phi_k(t)\rangle$ is periodic in time and that in accordance with the relation $E = \hbar\nu$, we write

$$|\phi_k(t)\rangle = |a_k\rangle e^{-i(\hbar\nu)t} \quad (30)$$

This form of the wave refers to a stationary state. The first expression of Eq. (27) now becomes

$$E|a_k\rangle = H|a_k\rangle \quad (31)$$

In matrix form, it can be written as

$$\begin{bmatrix} E & icp_3 & -icp_2 \\ -icp_3 & E & icp_1 \\ icp_2 & -icp_1 & E \end{bmatrix} \begin{bmatrix} a_{k1} \\ a_{k2} \\ a_{k3} \end{bmatrix} = 0 \quad (32)$$

Solutions for Eq. (32) require the vanishing of the determinant of $(E - H)$. In Eq. (32) the vector a_k is, of course, subject to the condition

$$\mathbf{k} \cdot \mathbf{a}_k = 0$$

By taking the determinant of $(E - H)$ we obtain

$$E(E^2 - c^2 p^2) = 0$$

The eigenvalue $E = 0$ implies [as seen from Eq. (32)] that the vectors a_k and p are parallel, which is not consistent with $p \cdot a_k = 0$. Hence the solution $E = 0$ must be discarded by the condition of transversality of the wave. The remaining two solutions are

$$E = cp = \hbar k = \hbar\nu$$

$$E = -cp = -\hbar k = -\hbar\nu$$

The negative sign in the second case is not to be understood as referring to a negative energy state. It has to do with the spin degree of freedom of the photon. This can be seen from writing the wave equation for negative energy,

$$(\mathbf{c} \cdot \mathbf{p}) |\eta\rangle = -E |\eta\rangle$$

and taking complex conjugate of both sides in the form

$$\mathbf{c} \cdot \mathbf{p} |\eta^*\rangle = E |\eta^*\rangle.$$

Thus both energy states are positive and they refer respectively to the energy of a right circularly or left circularly polarized photon, with spin parallel or antiparallel to \mathbf{k} . The two states can be represented by $|\eta\rangle$ and $|\eta^*\rangle$.

We have seen that at a given time, the value of the wave function can be obtained by a certain superposition of plane waves. In analogy to the classical definition of the intensity of a wave we shall look upon the real quantity

$$P = \langle \eta | \eta \rangle$$

as a probability density in the sense that EP is the energy density over a region that is large compared to the wavelength of the photon. The probability depends, of course, on the value of the wave function at the particular point. We can choose $|\eta\rangle$ in such a way that it differs from zero only over a region Ω of the dimensions of the wavelength of the photon. It shall be composed of monochromatic plane waves that interfere destructively outside the region Ω ; the frequencies and wavenumbers of these waves differ very little from each other inside the region Ω . In this particular region, the waves interfere constructively and the wave function $|\eta\rangle$ assumes large values there. This is equivalent to a localization of the photon. If this localization process is described in accordance with uncertainty relations between the dynamical variables of the photon, then the region in question is a wave packet. These arguments are, of course, valid also for mass particles.

The amplitudes of the waves that constitute a packet are different from zero only in the packet. These wave amplitudes constitute a wave group and have a group velocity that differs from the velocity of a single plane wave ("phase velocity").

In classical mechanics, the specification of positions and velocities of a dynamical system at each time is sufficient for a complete determination of its state. The basic concepts of classical theory (mechanics) consist of the material point, the force of interaction between material points (potential energy), and the inertial system (= the Cartesian coordinate system, the time coordinate or all reference frames in uniform relative motion). When the electromagnetic field is included, classical physics gains the concept of field, a region of space at every point of which a material particle can experience a force. Thus a field has an energy and momentum content. Special relativity brings into the structure of the inertial system the constancy of the velocity of light. In this theory, one cannot preserve the concepts of action at a distance and potential energy and this in turn implies that the concept of the material point can be discarded and replaced by the field concept.

A more profound change in our concepts of space and time came with the discovery of general relativity (the principle of general covariance,

invariance of the laws of nature with respect to all coordinate transformations, not necessarily in uniform relative motion). According to this theory, inertial systems need not be qualified as the only group of systems for the formulation of physical laws. The "space" represented by the inertial group (the Lorentz group), and considered as a part of the physical reality, can have only a limited meaning. The inertial group is used to determine the behavior of mass points in space and time, without itself being influenced by mass points. Therefore, according to special relativity, the inertial group occupies an absolute position in the description of physical phenomena.

In general relativity, however, the inertial group does not have this privileged position; in general relativity, it has been integrated into the field and has, therefore, been deprived of its "absoluteness." It is the field which has an independent meaning, and it depends on four parameters (the coordinates). The space aspect of matter itself is described by the field. The inertial character of real things must be derivable from a field; we do not try to fit a field to a given inertial group or to a mechanical system without considering that it will not change or influence the inertial system. In short, the mechanics of a system should be derivable from a field.

In quantum theory, we retain the inertial frame and the action-at-a-distance concepts, but it is no longer true that a state is completely defined by the initial conditions of the dynamical system.

According to quantum theory a "small system"—i.e., a system that can change its state (energy, momentum, position, angular momentum, and so on) by an act of observation—cannot be observed with the greatest possible amount of detail. The behavior of nature in the micro system is such that there exists a limitation or a lower limit to the power of observation. This limitation on the observability of a dynamical system implies a restriction on the data that can be assigned to the state of the physical system. Both classical and relativistic mechanics permit a definite distinction between observer and observed. This is essentially a complete fulfillment of the principle of causality in a deterministic sense. Accordingly, the interaction between observed and observer, arising in the act of observation, can be made infinitely small. For this reason it is meaningful, in classical mechanics, to say that the state of a dynamical system can be defined in its entirety with no limitation on the detail of the data. The continuous nature of things, for example the absorption of a wave by an electron, will allow enough time to measure its position and momentum rapidly before it absorbs sufficient energy from the incident wave to change its state abruptly.

The quantum mechanical point of view, at the expense of some loss of exact information on the dynamical system, recognized the impossibility of controlling the interaction between observer and observed. In the act of observation large changes in the state of the system being observed must be taking place. It is, therefore,

not possible to assign simultaneous "initial values" to canonically conjugate variables referring to the same degree of freedom of the dynamical system, e.g., x and p_x the coordinate and the corresponding linear momentum. This proposal of quantum mechanics can only be reconciled with a statistical or probabilistic approach to the description of physical reality. The complete determinism of classical theory is replaced by an indeterministic description, the extent of indeterminism being determined by the size of the universal constant h , Planck's constant. In classical mechanics, a cause causes another definite cause. In quantum mechanics, a cause can only produce a statistical trend to a given cause. This in turn can be regarded as a tendency towards some effect. The mode of the statistical tendency can be incorporated into a fundamental principle of nature, first enunciated by Heisenberg as the "principle of uncertainty" in the origination of physical events (see HEISENBERG UNCERTAINTY PRINCIPLE).

According to the principle of uncertainty, the position of an electron can be defined within a certain accuracy Δx at the time t , where Δx is a possible spread in the location of the electron; i.e., the electron can be seen within a region of the dimensions of Δx . The size of this region will depend on the spread of its momentum caused by the act of observation. In modern theory it is believed that the only description consistent with the principle of uncertainty is to represent the spread Δx as a "wave packet," the particle under observation being the wave itself. A wave packet has the properties of waves whose amplitudes are different from zero only in a limited region of space-time.

A dynamical system, having a well-defined single path in classical mechanics, is described in quantum mechanics by a wave packet containing coordinates and momenta with approximate numerical values that are restricted by the uncertainty principle. The size of the packet will increase with time; i.e., a wave packet can spread out and decay. If we succeed in obtaining, during our act of observation, a wave packet that can be located in a region smaller than Δx , then the location of the electron is sharper than ever. The latter result can only be approached as a result of large interaction between the electron and observer, resulting in a spread of the momentum of the wave packet much larger than the spread in its position. This uncertainty must be regarded as a fundamental property of a small object (electron) and, indeed, as its definition. The mathematical statement of the principle of uncertainty is contained in the relation

$$\Delta x \Delta p_x \geq h \quad (33)$$

The same kind of uncertainties will, of course, prevail for all other canonically conjugate dynamical variables of the wave packet—for example, the spread in its energy—and the corresponding spread in the time will satisfy the uncertainty relation

$$\Delta t \Delta E \geq h \quad (34)$$

The question now arises: what is it that the act of observation does to a dynamical system? What kind of a statement is the observer going to make following the observation of the dynamical system? An observable in quantum mechanics—such as energy, momentum, etc.—is, first of all, represented by a linear Hermitian operator. Let α be such an operator, with eigenstates $|\alpha'\rangle$ corresponding to its eigenvalue α' . If the state of the dynamical system during the act of measurement happens to be the eigenstate of α , then the eigenvalue α' is the result of the measurement. Whatever the state of the system prior to measurement was, the act of observation has caused it to jump to its eigenstate $|\alpha'\rangle$. A second measurement carried out in the state represented by $|\alpha'\rangle$ must give the same result. Therefore, any result of a measurement of an observable is one of its eigenvalues. The eigenvalues of a dynamical variable come into existence as a result of acts of measurement. Without the act of observation, we cannot talk of any state and consequently there exists no wave function prior to the process of measurement. Quantum mechanics is not concerned with the state and the corresponding possible values of a dynamical variable prior to observation. Quantum mechanics predicts the future from the present data

in accordance with the principle of uncertainty. In general, every experiment aimed at a determination of some numerical quantity causes a loss of information in some other quantity, related to the former by uncertainty relations. The uncontrollable perturbation of the observed systems alters the value of the previously determined quantities, except when the corresponding linear Hermitian operators commute with one another, i.e., the measurement of one is not affected by the measurement of the other.

BEHRAM KURSUNOGLU

References

- Bethe, H. A., "Intermediate Quantum Mechanics," New York and Amsterdam, W. A. Benjamin, Inc., 1964.
 Dirac, P. A. M., "The Principles of Quantum Mechanics," Fourth edition, Oxford, Clarendon Press, 1958.
 Kurşunoğlu, B., "Modern Quantum Theory," San Francisco and London, Freeman and Co., 1962.

Cross-references: HEISENBERG UNCERTAINTY PRINCIPLE; MATRIX MECHANICS; PHOTON; RADIATION, THERMAL; SCHRÖDINGER EQUATION; WAVE MECHANICS.

R

RADAR

Radar is the name given to the use of electromagnetic energy for the detection and location of reflecting objects. It operates by transmitting an electromagnetic signal and comparing the echo reflected from the target with the transmitted signal.

The first demonstration of the basic radar effects was by Heinrich Hertz in his famous experiments in the late 1880's in which he verified Maxwell's electromagnetic theory. Hertz showed that short-wave radiation could be reflected from metallic and dielectric bodies. Although the basic principle of radar was embodied in Hertz's experiments, the practical development of radar had to wait more than 50 years for radio technology to advance sufficiently. It wasn't until the late 1930's that practical models of radars appeared. The rapid advance in radar technology during World War II was aided by the many significant contributions of physicists and other scientists pressed into the practical pursuit of a new technology important to the military. In addition to its military application, radar has been applied to the peace-time needs of air and ship navigation, air traffic control, rainfall observation, tornado detection, hurricane tracking, surveying, radar astronomy, and the familiar speed measuring meter of the highway police (see PROPAGATION OF ELECTROMAGNETIC WAVES).

The measurement of distance, or range, is probably the most distinctive feature of radar. Range is determined from the time taken by the transmitted signal to travel out to the target and back. The distances involved might be as short as a few feet or as long as interplanetary distances.

If the target is in motion relative to the radar, the echo signal will be shifted in frequency by the DOPPLER EFFECT and may be used as a direct measurement of the relative target velocity. A more important application of the doppler shift is to separate moving targets from stationary targets (clutter) by means of frequency filtering. This is the basis of MTI (moving target indication) radar.

Radar antennas are large compared to the wavelength so as to produce narrow, directive beams. The direction of the target may be inferred from the angle of arrival of the echo. Radar antenna technology has profited greatly from the theory and practice of optics. Both the lens and

the parabolic mirror have their counterpart in radar, and the analysis of antenna radiation patterns follows from diffraction theory developed for optics. The greater versatility of materials in the radar frequency region of the electromagnetic spectrum, however, offers more flexibility in implementing many of the principles of optics not practical in the visual portion of the spectrum.

The external appearance of a radar is dominated by its ANTENNA. Most radars use some form of parabolic reflector. The radar antenna can also be a fixed array of many small radiating elements (perhaps several thousand or more) operating in unison to produce the desired radiation characteristics. Array antennas have the advantage of greater flexibility and more rapid beam steering than mechanically steered reflector antennas because the beam movement can be accomplished by electrically changing the relative phase at each antenna element. High power can be radiated since a separate transmitter can be applied at each element. The flexibility and speed of an array antenna make it necessary in many instances to control its functions and analyze its output by automatic data processing equipment rather than with an operator using a grease pencil to mark the face of a cathode ray tube.

Radars are generally found within the microwave portion of the electromagnetic spectrum, typically from about 200 Mc (1.5 meters wavelength) to about 35,000 Mc (8.5 mm wavelength). These are not firm bounds. Many radars operate outside these limits. The famous British CH radar system of World War II which provided warning of air attack, operated in the high-frequency region of the spectrum in the vicinity of 25 Mc. Experimental radars have also been demonstrated in the millimeter wavelength region where small physical apertures are capable of narrow beam widths and good angular resolution. The radar principle has also been applied at optical frequencies with LASERS for the measurement of range and detection of small motions (using the Doppler effect).

The detection performance of a radar system is specified by the *radar equation* which states

$$P_{\text{rec}} = \frac{P_t G}{4\pi R^2} \times \sigma \times \frac{1}{4\pi R^2} \times A$$

$$\text{Received power} = \frac{P_t \times G \times \sigma \times A}{R^4}$$

Power Target Space Antenna
 density back- atten- collec-
 at a scatter uation ting
 distance cross on area
 section return
 path

where P_t is the transmitted power, G is the transmitting antenna gain, R is the range, σ is the backscatter cross section, and A is the effective receiving aperture of the antenna. The wavelength λ of the radar signal does not appear explicitly in this expression, but it can be introduced by the relationship between the gain and effective receiving area of an antenna which states $G = 4\pi A/\lambda^2$.

The detection capability and the measurement accuracy of a radar are ultimately limited by noise. The noise may be generated within the radar receiver itself, or it may be external and enter the receiver via the antenna along with the desired signal. External noise is generally small at microwave frequencies, but it can be a significant part of the overall noise if low-noise receiving devices such as the MASER and the parametric amplifier are used (see MEASUREMENTS, PRINCIPLES OF).

Since the effects of noise must be considered in statistical terms, the analysis and understanding of the basic properties of radar have benefited from the application of the mathematical theory of statistics. The statistical theory of hypothesis testing has been applied to the radar detection problem where it is necessary to determine which of two hypotheses is correct: the output of a radar receiver is due to (1) noise alone or (2) signal plus noise. One of the results is the quantitative specification of the signal-to-noise ratio required at the receiver for reliable detection. Also derived from hypothesis testing based on the likelihood ratio or a *posteriori* probability are concepts for ideal detection methods with which to compare the performance of practical receivers. The statistical theory of parameter estimation has also been applied with success to analyze the accuracy and theoretical limits of radar measurements.

Reliable detection of targets requires signal-to-noise power ratios of the order of 10 to 100 at the receiver, depending on the degree of error that can be tolerated in making the decision as to the presence or absence of a target. Even larger values are generally needed for the accurate measurement of target parameters. (These values may seem surprisingly high but, for comparison, the minimum signal-to-noise ratio of quality television signals is usually of the order of 10,000.)

The rms error δT in measuring the time delay to the target and back (the range measurement) can be expressed as

$$\delta T = \frac{1}{\beta \sqrt{2EN_0}}$$

where β is defined as the effective signal bandwidth, E is the total energy of the received

signal, and N_0 is the noise power per unit cycle of bandwidth assuming the noise has a uniform spectrum over the bandwidth of the receiver. The square of β is equal to $(2\pi)^2$ times the second central moment of the power spectrum normalized with respect to the signal energy. For a simple rectangular pulse, E/N_0 is approximately equal to the signal-to-noise (power) ratio. To obtain an accurate range measurement, E/N_0 and the signal bandwidth must be large. A similar expression applies to the accuracy of the measurement of Doppler frequency if the rms time delay error is replaced by the rms frequency error and the effective bandwidth is replaced by the effective time duration of the signal. Thus, the longer the signal duration and the greater the ratio E/N_0 , the more accurate is the Doppler frequency measurement. Likewise the angular measurement accuracy also depends on the ratio E/N_0 and the effective aperture size.

In addition to noise, radar can be limited by the presence of unwanted interfering echoes from large nearby objects such as the surface of the ground, trees, vegetation, sea waves, and weather. Although these "clutter" echoes may be troublesome in some applications, they are sometimes the echoes of interest as, for example, in ground mapping and meteorological applications.

Radar comes in many sizes and shapes. The smallest can be held in the hand and might radiate as little as a few milliwatts and be used to detect the movement of people at short range. The largest might radiate megawatts of average power and operate with antennas the size of a football field and would be used for the detection of space objects at ranges of many thousands of miles or more.

MERRILL I. SKOLNIK

References

- Ridenour, L. N., "Radar System Engineering," MIT Radiation Laboratory Series, Vol. 1, New York, McGraw Hill Book Co., 1947.
- Woodward, P. M., "Probability and Information Theory, with Applications to Radar," New York, McGraw Hill Book Co., 1953.
- Battan, L. J., "Radar Meteorology," Chicago, University of Chicago Press, 1959.
- Skolnik, M. I., "Introduction to Radar Systems," New York, McGraw-Hill Book Co., 1962.
- Carpentier, M., "Radars Theories Modernes," Dunod, Paris, 1963.

Cross-references: ANTENNAS; DOPPLER EFFECT; LASER; MASER; MEASUREMENTS, PRINCIPLES OF; MICROWAVE TRANSMISSION; PROPAGATION OF ELECTROMAGNETIC WAVES.

RADIATION BELTS

The first U.S. satellite Explorer 1 in 1958 carried Geiger counters designed by Van Allen of Iowa. This instrument discovered the high flux of energetic charged particles trapped in the

magnetic field of the earth now called the Van Allen radiation belt. We now know that this belt is made up of high-energy protons and electrons put into the magnetic field from several different sources. The particles travel in helical paths along field lines bouncing back and forth from one hemisphere to the other. The fact that the field lines converge towards the earth produces a force on the particle of $F = evB_{\perp}$ where B_{\perp} is the component of the magnetic field perpendicular to the field lines. This force pushes the particle out of the region of converging field and produces the bouncing motion. Without this bouncing motion, radiation belts would not exist. Other forces such as the centrifugal force on a particle moving along a curved field line make the particle drift in longitude around the earth. This combination of motions makes a roughly spherical belt of particles around the earth. The particles are not observed below about 500 km altitude. The fact that the earth's magnetic field is not symmetric but is weaker in the South Atlantic means the trapped particles are observed at lower altitudes there. The absence of trapped particles at low altitudes is clearly due to the presence of the atmosphere. The particles are lost very rapidly below this by coulomb scattering and energy loss to the thermal atmospheric atoms.

In 1959, Freden and White flew nuclear emulsions on an Atlas rocket, recovered them, and measured the tracks to show that the most energetic particles in the inner part of the radiation belt were protons with energies up to at least 700 MeV. This particle population stays constant for long periods of time with a maximum flux of about $10^4 \text{ cm}^{-2} \text{ sec}^{-1}$. We understand well how these high-energy protons are made. Very high-energy cosmic rays from space bombard the earth's atmosphere, collide with nitrogen or oxygen nuclei, and produce neutrons. Some of these neutrons emerge from the top of the atmosphere and, being radioactive, decay in space, producing protons and electrons. A quantitative study of these neutron-decay protons shows they have all the right properties to be the source of the observed protons from 10 MeV up.

The radiation belt extends out to about 10 earth radii and then stops abruptly. This outer edge is caused by the action of the sun on the earth. The solar corona is very hot and is unstable. Because of this, the sun continuously blows a wind of protons and electrons at the earth with a velocity of about 400 km/sec. This plasma pushes in the geomagnetic field until the solar wind pressure is balanced by $B^2/8\pi$, the pressure of the distorted geomagnetic field pushing outwards. Thus a boundary, called the magnetopause, exists between the solar environment and the geomagnetic field. Only inside the cavity, called the magnetosphere, occupied by the earth's field can trapped particles exist. Outside this cavity, a zone of energetic electrons exists, apparently made as a result of the solar wind hitting the magnetosphere and producing a turbulent region. These particles are not trapped.

The low energy trapped electrons found inside the magnetosphere have similar energies and flux to those outside, suggesting that the boundary is leaky and the electrons somehow diffuse in across the boundary.

We know that auroras are made mostly by energetic electrons striking the upper atmosphere. There is clearly a relationship between auroras and the trapped electrons found in the outer part of the radiation belt. They seem to be made by the same process (see AURORA AND AIRGLOW).

A large population of protons of about $10^6 \text{ cm}^{-2} \text{ sec}^{-1}$ is found in the outer part of the magnetosphere. They have energies of the order of 1 MeV. The protons have systematically higher energies closer to the earth as would be expected if they were drifting inwards and being accelerated by the increasing magnetic field as they go. A varying intensity solar wind can change the size of the magnetosphere, and as a result, a series of changes magnetically pump these particles inward to give the observed properties of the protons.

Clearly, all the processes operating to introduce particles into the radiation belt or all the ways in which particles are lost are not understood. All of the components of the trapped radiation have probably not even been found because we have not studied low-energy (less than 100 eV) particles well yet. One of the most important fields of study of the future will be to determine the types and occurrence rates of various natural magnetic disturbances at high altitudes which very likely play an important part in perturbing particle motion. A considerable amount has been learned about the natural Van Allen belt from studying charged particles artificially injected into the magnetosphere.

On seven occasions since 1958, artificial radiation belts have been made by the explosion of nuclear bombs at high altitude. The belt results from the β decay of the fission fragments made in the explosion. The decay produces several electrons per fission fragment with energies averaging about 1 MeV but extending up to 8 MeV. In 1958, three nuclear explosions in the South Atlantic, called the Argus experiment, were carried out to show that the earth's field could store charged particles. The planning for Argus was well along before Van Allen discovered the natural radiation belt. These explosions did make artificial radiation belts which were studied by the Explorer IV satellite.

In July 1962, the U.S. Starfish explosion of a 1.4-megaton bomb at 400 km over the central Pacific made a large artificial belt with electron fluxes up to $10^6 \text{ cm}^{-2} \text{ sec}^{-1}$. The satellites Injun I and III, Telstar I and Explorer XV followed the belt decay and showed that there are two different processes acting to remove the electrons. Below about 3000 km altitude, the electrons are lost mostly by coulomb scattering on the nuclei of the air atoms present, changing their direction of motion and diffusing down magnetic field lines into the atmosphere. An electron at 2000 km lives roughly a year before being lost this way.

At higher altitudes, the electrons are lost faster than this. Three Soviet high altitude explosions in the fall of 1962 produced artificial belts that decayed similarly to the Starfish belt. At 10,000 km altitude, the electron lifetime is just a few days. It seems probable that the electrons are lost here by interacting with electromagnetic waves, called whistlers. These waves are circularly polarized waves traveling along field lines. When the particle's gyration frequency resonates with the whistler wave frequency, the interaction can disturb the particle's motion and scatter it out of the trapping region.

We have some knowledge of radiation belts on other planets besides the earth. The Mariner space craft went within 40,000 km of Venus. It did not find a measurable planetary magnetic field or any trapped radiation belt. This does not necessarily mean that Venus does not have one at all. It does mean that if a magnetosphere filled with trapped particles exists on Venus, it must be quite small in size. Radio waves received from Jupiter, identified as synchrotron radiation from electrons gyrating around magnetic field lines, show that Jupiter has a trapped radiation belt. It must have considerably more high-energy trapped electrons than the earth's belt has because we haven't been able to observe synchrotron radiation from the natural Van Allen belt.

WILMOT N. HESS

Cross-references: ELECTRON, GEOPHYSICS, IONOSPHERE, PLANETARY ATMOSPHERES, PROTON, SPACE PHYSICS.

RADIATION CHEMISTRY

Radiation chemistry is the study of the chemical effects produced by the absorption of ionizing radiation. It includes chemical effects produced by the absorption of radiation from radiative nuclei (α , β , and γ rays), of high-energy charged particles (electrons, protons, recoil nuclei, etc.), and of electromagnetic radiation of short wavelength [x-rays with a wavelength less than about 100 Å and an energy greater than about 100 eV, for example]. Electromagnetic radiation of rather longer wavelength, in the ultraviolet and visible regions of the spectrum, may also initiate chemical reactions, though normally without producing ions as reactive intermediates; such reactions are the province of photochemistry. Reactions chemically similar to those caused by the absorption of ionizing radiation can be initiated by electric discharges; others, initiated in various ways, occur in the upper atmosphere.

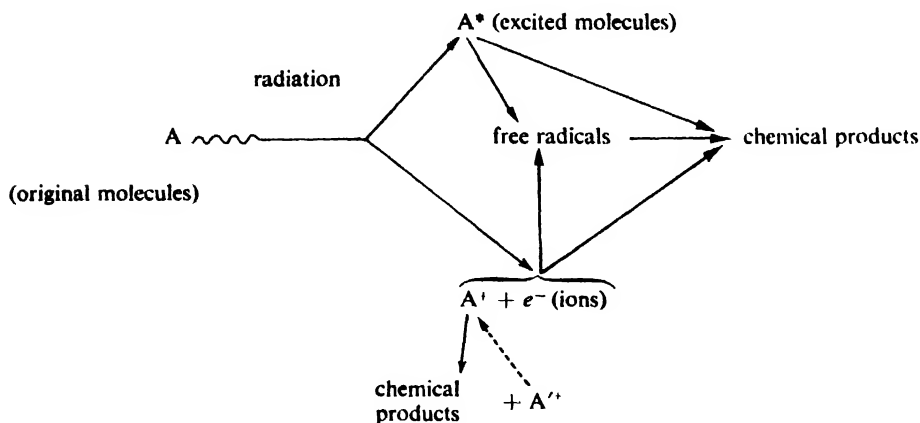
Radiation chemistry originated with the observations by Röntgen (1895) and by Becquerel (1896) that led to the discovery of x-rays and of radioactivity, namely that photographic plates become fogged when placed near discharge tubes and uranium salts respectively. The subject was studied to only a limited extent until about 1942, when the advent of nuclear reactors and the increased interest in high-energy physics provided both the incentive and the means (in the

form of relatively cheap artificial radioactive isotopes and of large particle accelerators) for a more intensive study. In the following two decades, earlier work was consolidated and the basic mechanisms for radiation-induced action were established in outline, making possible an extremely rapid development of the subject since about 1960.

Absorption of any form of ionizing radiation by matter produces positive ions (A^+), electrons (e^-), and electronically excited atoms or molecules (A^*) distributed along the tracks of charged particles. The charged particles may be those which comprise the radiation (e.g., electrons or helium nuclei with fast electron or α -irradiation) or secondary particles produced by interactions of the primary radiation (e.g., fast secondary electrons formed by the absorption of x- or γ -rays, or fast protons produced by interaction of neutrons with hydrogenous materials). In addition, some of the electrons produced by ionization in the medium will have sufficient energy to produce further ionization and excitation, and will do so in slowing down to thermal energy (such electrons are known as δ rays). Except in gases, where the ions and excited molecules can diffuse apart quite readily, these primary species are initially concentrated within about 10 Å of the track of the ionizing particle. Heavy charged particles (protons and, particularly, helium nuclei and heavier particles) lose energy very rapidly in liquids and solids and leave a track densely populated with the primary species; radiations of this type are said to have a "high LET," where LET stands for linear energy transfer. In contrast, fast electrons and the secondary electrons produced by the absorption of x- and γ rays lose energy relatively slowly and form the primary species (which are the same as those produced by the heavier particles) in small groups containing, on an average, two or three ion pairs and about the same number of excited molecules. These groups (called *spurs*), with an initial diameter of ~ 20 Å, are separated by a relatively great distance ($\sim 10^3$ Å) from neighboring spurs along the same track, and the radiation responsible is said to have a "low LET." The initial localization of the ions and excited molecules in the track of a high-LET particle or in spurs produced by a low-LET particle causes spatial, or track, effects in radiation-induced reactions which are absent in photochemistry, where excited molecules are produced with an essentially uniform distribution in any plane at right angles to the direction of the beam of light. The excited molecules involved in radiation chemistry include excited states similar to those formed by the absorption of ultraviolet or visible light and also other states, formed by optically forbidden transitions or with more intrinsic energy, that are not produced photochemically.

Chemical changes in the irradiated material are brought about by breakdown or reaction of the ions and excited molecules and via free radicals formed by these primary species. Free radicals are chemically very reactive and are

produced in almost all materials upon irradiation. Their study has been greatly aided by the introduction of electron spin resonance (ESR) and pulsed radiation techniques. The sequence of events in many radiation-induced (radiolysis) reactions may be represented as:



The arrows represent dissociation of the intermediates, reaction between these species, and reaction with molecules of the substrate.

Quantitative studies in radiation chemistry are based either on the number of ions (N) formed in the irradiated medium or on the energy absorbed from the radiation. Ionization measurements in irradiated gases allow N to be estimated, and radiation yields in gases are best expressed as the ionic (or *ion-pair*) yield, M/N , where M is the number of molecules of product formed by radiation which produces N ion pairs. Reliable estimates of N are, at present, only available for gaseous systems, and for condensed materials it is usual to express the radiolysis yield as a G value (the number of molecules of product formed, or of starting material changed, per 100 eV of energy absorbed). Ionic yields and G values are related by $G = M/N \times 100/W$, where W (electron volts) is the mean energy required to form an ion pair in the material being irradiated. The determination of N , or of the energy absorbed, is termed "dosimetry" and may involve ionization measurements, calorimetry, measurement of the charge carried by a beam of charged particles of known energy, or measurement of a chemical change produced by irradiation [the oxidation of ferrous iron to ferric iron in 0.4 M sulfuric acid solution is often used (Fricke dosimeter)]. However, chemical dosimeters must first be calibrated against some absolute physical measurement.

Radiation-induced reactions have been studied in the gas, liquid and solid phases and in inorganic, organic, and biological systems. They have also been studied over a wide range of temperatures.

Most thoroughly studied of all radiation-induced reactions is the radiolysis of water and aqueous solutions, where the following have been

identified as the major products formed in the tracks or spurs: hydrogen (H_2), hydrogen peroxide (H_2O_2), hydrogen atoms ($H\cdot$), hydroxyl radicals ($\cdot OH$), and solvated, or hydrated, electrons (e_{aq}^-) (the formation of solvated elec-

trons is probably typical of polar liquids). These products are thought to be formed by the primary species, H_2O^+ , e^- , and H_2O^* , and, in the case of hydrogen and hydrogen peroxide, by $H\cdot$, $\cdot OH$, and e_{aq}^- . Relatively more hydrogen and hydrogen peroxide are formed in the track of a high-LET α -particle [$G(H_2) = 1.57$, $G(H_2O_2) = 1.45$] than in the spurs associated with a high-energy, low-LET, electron [$G(H_2) = 0.40$, $G(H_2O_2) = 0.80$] since the concentration of the precursors is greater in the track than in the spurs. However, relatively more of the other products ($H\cdot$, $\cdot OH$, and e_{aq}^-) escape from the spurs than from the α -particle track. These three species are free radicals and react very readily with substances present in solution, the hydroxyl radical generally producing oxidation, and the hydrogen atom and the solvated electron producing reduction, of the solute if the solution is free of air. Both the hydrogen atom and the solvated electron react rapidly with oxygen to give oxidizing species, so that oxidation reactions predominate when aerated aqueous solutions are irradiated.

Many organic materials have been irradiated and, making a very rough generalization, the products are those expected if the action of the radiation is to break the organic molecules randomly into two fragments (free radicals). The fragments then react together in pairs, again in a random fashion, either combining to form larger molecules or transferring an atom from one fragment to the other to give two stable molecules. The products from a hydrocarbon, for example, include hydrogen and hydrocarbons ranging in size from methane (CH_4) to compounds containing twice as many carbon atoms as the original compound; unsaturated materials, containing relatively less hydrogen than the original compound, will also be formed. Hexane,

$\text{CH}_3(\text{CH}_2)_4\text{CH}_3$, forms at least 16 products upon irradiation, of which the two most abundant are hydrogen ($G \approx 5.0$) and the dimeric C_{12} hydrocarbon ($G \approx 2.0$); the remaining products are formed with G values of 0.5 or less. More complex compounds may break preferentially (though generally not exclusively) at a particular point in the molecule; thus the carbon-iodine bond breaks most frequently when methyl iodide CH_3I , is irradiated. Some classes of compound can enter into chain reactions in which the reaction, once started, continues on its own. Typical of such substances are the unsaturated "monomers" which polymerize to produce polymers such as polymethyl methacrylate ("Lucite") and polyvinyl chloride (PVC). The chain reactions here are initiated by free radicals, and other means of producing free radicals besides irradiation bring about the same effect. G values for chain reactions may run from a few hundred to many thousands.

Aromatic compounds such as benzene are more resistant to radiation than most compounds lacking the "benzene" ring. Thus $G(\text{H}_2)$ from liquid benzene is only 0.036, and the total yield of radiolysis products does not exceed $G \approx 1$. In favorable instances, aromatic compounds can reduce the radiation damage in non-aromatic compounds mixed with them, energy absorbed by the second component being transferred in part to the "protecting" aromatic compound.

Apart from its intrinsic interest, the importance of radiation chemistry until the present time has rested largely upon its application to problems of reactor technology and its close relationship to radiation biology and radiation medicine. Recently, however, several industrial applications have been realized, e.g., the treatment of polyethylene to produce a higher-melting polymer and the synthesis of ethyl bromide via a radiation-induced chain reaction between ethylene and hydrogen bromide. Other commercial applications will undoubtedly follow, probably depending for their success on the special properties of high-energy radiation since it seems unlikely that high-energy radiation will ever be a very cheap means of introducing energy into a chemical system.

J. W. T. SPINKS
R. J. WOODS

References

- Spinks, J. W. T., and Woods, R. J., "An Introduction to Radiation Chemistry," New York, John Wiley & Sons, 1964.
Hart, E. J., and Platzman, R. L., in Herrera, M., and Forssberg, A., Eds., "Mechanisms in Radiobiology," Vol. 1, Ch. 2, New York, Academic Press, 1961.
Haissinsky, M., Ed., "Actions Chimiques et Biologiques des Radiations," Vol. 1, Paris, Masson et Cie., 1955, and succeeding volumes.
Review articles in *Ann. Rev. Phys. Chem.*, 1 (1950) and succeeding volumes.

- Hine, G. J., and Brownell, G. L., Eds., "Radiation Dosimetry," New York, Academic Press, 1956.
Lind, S. C., "Radiation Chemistry of Gases," New York, Reinhold Publishing Corp. 1961.
Allen, A. O., "The Radiation Chemistry of Water and Aqueous Solutions," Princeton, N. J., D. Van Nostrand, 1961.
Swallow, A. J., "Radiation Chemistry of Organic Compounds," New York, Pergamon Press, 1960.
Charlesby, A., "Atomic Radiation and Polymers," New York, Pergamon Press, 1960.
Chapiro, A., "Radiation Chemistry of Polymeric Systems," New York, Interscience Publishers, 1962.
Brownell, L. E., "Radiation Uses in Industry and Science," U.S. Atomic Energy Commission, 1961.

Cross-references: IONIZATION; PHOTOCHEMISTRY; RADIATION, IONIZING, BASIC INTERACTIONS; SECONDARY EMISSION.

RADIATION, IONIZING, BASIC INTERACTIONS

Radiations of a large class, called ionizing, which interact with matter in its many forms and lead to a wide variety of "observed" or expressed effects have similar basic interaction pathways. Sources of ionizing radiations are varied and include radioactive isotopes, fission and fusion reactions, particle accelerators, and cosmic rays. Regardless of the source of any given radiation (applied to a given target), the basic interaction depends only upon the fundamental properties of the radiation itself.

Ionizing Radiations. The principal ionizing radiations are summarized in Table I. Although only the gamma or x-rays are electromagnetic in character and thus "radiations" in the classical sense, the distinction between "radiations" and ionizing "particles" is often not made (x-rays are distinguished from gamma rays only with respect to their origins; gamma rays result from nuclear interactions or decays; x-rays result from transitions of atomic or free electrons, produced artificially by bombarding metallic targets with energetic electrons). It is sometimes difficult to make a clear distinction between ionizing and nonionizing electromagnetic radiations, particularly in condensed phases. The ionization potential of *gaseous* elements i.e., the energy required for removal of the first electron, varies from 3.9 eV (Cs) to 24.6 eV (He). Comparable values are not well known for most complicated molecular systems or liquid- and solid-state systems, but they are probably in or near this general range. Although ultraviolet and even visible light can in special cases cause ionizations, the general assumption is that more energetic x- or gamma radiation is required to insure ionization. Hence the name "ionizing radiation" is reserved for electromagnetic radiation at least as energetic as x-rays and for charged particles of similar energies. Neutrons also lead to ionization, but for other reasons, described below. *Ionization* is, of course, not the sole interaction of high-energy

TABLE I

Name	Symbol	Location in Atom	Relative Rest Mass	Charge
Proton (H ¹) ⁺	<i>p</i>	Nucleus	1	+1
Neutron	<i>n</i>	Nucleus	1	0
Electron	<i>e</i>	Outer shells	0.00055	-1
Beta ⁻ (electron)	β^-	Emitted during decay processes	0.00055	-1
Beta ⁺ (positron)	β^+		0.00055	+1
Alpha (He ⁴) ⁺⁺	α		4	+2
Gamma* (photon)	γ	Emitted during decay processes	0.0	0

* X-rays of equal energy are identical, but of extranuclear origin.

particles and radiations with matter. The *excitation* of atomic electrons into higher-energy states always accompanies ionization.

Basic Action. The fundamental processes leading to ionization differ for charged particles, electromagnetic radiations, and neutrons.

Charged Particles. Fast charged particles are produced in radioactive decay processes, particle accelerators, nuclear reactions, and extraterrestrial sources. They undergo *coulombic* and *nuclear* interactions. The latter are far less probable and are discussed in other parts of this encyclopaedia (see NUCLEAR REACTIONS). There are two principal means whereby charged particles can lose energy by coulomb interaction: radiative loss and direct ionization. The probability of radiative energy loss (BREMSSTRAHLUNG) is roughly proportional to

$$\frac{z^2 Z^2 T}{M_0^2}$$

where *z* is the particle charge in units of the electron charge *e*, *Z* is the atomic number of the target material, *T* is the particle kinetic energy,

and *M₀* is the rest mass of the particle. The ratio of energy lost by bremsstrahlung to that by ionization can be approximated by

$$\left(\frac{m_0}{M_0}\right)^2 \frac{ZT}{1600 m_0 c^2}$$

in which *m₀* is the rest mass of the electron and *c²* is the speed of light squared, or 931 MeV per atomic mass unit. Electrons in the 10-MeV region lose about half of their energy by bremsstrahlung (in high-*Z* material), whereas heavier charged particles lose nearly all of their energy by ionization.

Loss of energy by ionization results when the particle undergoes coulomb collision with the *electrons* of the target. From the ratio of total energy lost to the number of ion pairs produced in cloud chambers (see Fig. 1), it has been estimated that approximately 100 eV are dissipated in each *primary ionization event* and that each event results in the production of, on the average, three ion pairs, each consisting of a free electron and a positive ion. If a resulting electron carries

IONIZATION BY DELTA RAYS



FIG. 1. Schematic drawing of a cloud-chamber photograph of the track of an ionizing particle, illustrating that ion pairs (illustrated by small circles) occur in clusters. Spurs indicated by arrows are ionizations due to *delta rays*. The arbitrary identification of the *track core* is also indicated.

more than a certain amount (usually considered to be about 100 eV) of kinetic energy, it is called a *delta ray*, because it is capable of further ionizations. "Delta ray" and "track core" ionizations are depicted in Fig. 1. The experimentally determined energy required to produce each *observed* ion pair lies in the range of 20 to 40 eV per ion pair for gases. A figure in the neighborhood of 30 eV per ion pair is commonly assumed for condensed phases of matter. This value is called "*W*" and is generally slightly different for electrons and heavy particles in the same material.

On the basis of coulomb scattering theory, it is possible to calculate the energy loss per unit path length ($-dT/dx$) for charged particles. For electrons and positrons the theory is complicated by the quantum mechanical effects of spin, identity, and relativity. In general, the *total* expectation energy loss per unit path length (also termed *mass stopping power* when path length is expressed in grams per square centimeter) for electrons and positrons can be expressed by

$$-\frac{dT}{dx} = \frac{2\pi e^4}{m_0 V^2} NZ \left\{ \ln \left[\frac{m_0 V^2 T}{2I^2(1 - \beta^2)} \right] - \beta^2 \right\}$$

in which V is the particle velocity, β is V/c , N is the number of atoms per unit mass, and I is the geometric mean ionization and excitation energy of the atoms of the target material. Classical theory is more adequate in describing the loss of energy by heavy charged particles, and the mass stopping power is determined from

$$-\frac{dT}{dx} = \frac{4\pi e^4 z^2}{m_0 V^2} NB$$

where B is called the *stopping number* and assumes various forms, some examples of which are

$$B = Z \ln \frac{2m_0 V^2}{I}$$

for nonrelativistic particles in material of low Z ;

$$B = Z \left[\ln \frac{2m_0 V^2}{I} - \ln(1 - \beta^2) - \beta^2 \right]$$

for relativistic particles in materials of low Z ; and

$$B = Z \left[\ln \frac{2m_0 V^2}{I} - \ln(1 - \beta^2) - \beta^2 - \frac{C_K}{Z} \right]$$

for relativistic particles in material of high Z , in which C_K is a term which corrects for the relative unavailability of K electrons for coulomb interactions. Corrections for L and M electrons may be required for very high- Z material.

Instead of stopping power, the term *linear energy transfer* (LET) is frequently used. The two quantities are not strictly interconvertible, according to a recent decision by the international committee on radiation units (see *National Bureau of Standards (U.S.) Handbook 84*). Mass stopping power should be used only in the sense

described above, whereas LET should be used only for energy lost by the particle within a specified distance of the track core. (Stopping power is usually expressed in units of million electron volts per gram per square centimeter, and LET in units of kiloelectron volts per micron.) These are both measures of linear *density of ionization* in the target material.

From the above equations, it is possible to derive range-energy relationships for charged particles by simple integration. Theoretical and experimental range-energy curves are now available in the literature for nearly all charged particles.

Electromagnetic Radiation (X, gamma). For fast, charged particles interacting with matter we had, above, relations of the form

$$-\frac{dT}{dx} \propto \frac{(\text{charge})^2}{(\text{velocity})^2}$$

If it were permissible to extend this idea to electromagnetic radiation, one might expect photons with no charge (and with the velocity of light) not to ionize at all. They indeed do ionize sparsely though for different reasons. The three principal mechanisms of interaction can be summarized as follows:

(1) Photoelectric effect. Low-energy photons can give up all their energy to a bound electron, forming an ion pair and disappearing in the process. (Generally unimportant above 1 MeV.)

(2) Compton scattering. For medium-energy photons (0.5 to 5 MeV), this elastic collision process predominates, leading to ejection of a recoil electron plus the partially degraded (longer-wavelength) photon.

(3) Pair production. Photons of highest energy most often interact by forming an electron-positron pair in the field of a nucleus and disappearing in the process. The absolute energy threshold for this process is the rest mass energy of the pair: 1.02 MeV.

The net result of all three processes is the formation of (charged) ion pairs and in particular of electrons having energies ranging up to the photon energy but on the average only a fraction of this maximum. The discussion on charged particles thus applies to x- or gamma-ray action also. Each of these processes will now be considered in more detail.

In the *photoelectric process*, the photon collides directly with an atomic electron, imparting kinetic energy

$$T = h\nu - B_e$$

where B_e is the atomic binding energy of the electron and $h\nu$ is the photon energy. For a considerable range of photon energies and target material Z 's, the probability of this process is approximately dependent upon the fourth power of Z and the inverse third power of the photon energy. There are, however, very specific energies at which photoelectric absorption is very probable ("absorption edges"), due to nearness of the photon energy to the binding energy of a

specific electron (K, L, M, etc.). The probability per unit thickness of absorber of a photon undergoing photoelectric absorption is denoted by τ . An initial photon intensity I_0 is reduced by this process to an intensity I_x after traversing a thickness of material, x , according to the relation

$$I_x/I_0 = e^{-\tau x}$$

The *Compton process* differs from the photoelectric process in two important ways: the photon usually loses less than all of its energy to the electron with which it collides, giving rise to a lower-energy scattered photon, and the process may occur with a free or loosely bound electron. Momentum and energy are conserved according to the laws of classical mechanics, and the energetics can be described by

$$h\nu = h\nu' + (m - m_0)c^2$$

where $h\nu'$ is the energy of the secondary photon, and $(m - m_0)c^2$ is the kinetic energy of the recoil electron (expressed in terms of the relativistic mass increase). At high photon energies, the collision probabilities are governed by the laws of quantum mechanics. Photons are lost by the Compton process according to the relation

$$I_x/I_0 = e^{-\sigma x}$$

where σ is the absorption probability per unit thickness.

If the energy of the photon is greater than two electron masses (1.02 MeV), there is a finite probability that it will interact with the nuclear field giving rise to electron-positron *pair production*. In pair production, energy is conserved according to

$$h\nu - 2m_0c^2 = T_+ + T_-$$

where T_+ and T_- are the positron and electron kinetic energies, respectively. The positive electron is ultimately annihilated by a negative electron, and the masses of both are converted into photons with energies distributed about 0.51 MeV (one electron mass each). Very high-energy electron-positron pairs may, in turn, produce further photons by bremsstrahlung, initiating a sequence of photon-electron-photon, etc., interactions, known as a *cascade*. (If the photon energy is greater than four electron masses, pair production may occur in the field of an electron, under which condition the original electron is set in motion, and the process is called "triplet production.") The probability of pair production increases very rapidly with increasing photon energy above 1.02 MeV and increases as Z^2 , the square of the atomic number of the target material. The absorption probability per unit thickness is denoted by κ , so that high-energy photons are absorbed according to

$$I_x/I_0 = e^{-\kappa x}$$

It is usually desirable to know the *total* photon absorption per unit thickness of a given material,

so one simply states

$$I_x/I_0 = e^{-\mu x}$$

where

$$\mu = \tau + \sigma + \kappa.$$

μ is called the "total linear absorption coefficient" and is expressed in units reciprocal to those in which x is measured.

Neutrons. Because of their lack of charge, neutrons do not interact electrostatically either with orbital electrons or with nuclei. They do interact with nuclei, however, in various other ways.

Fast neutrons (up to a few MeV) lose energy primarily by elastic collisions with other nuclei. From considerations of momentum transfer, this process is most efficient for target nuclei of about the same mass (i.e., protons in hydrogenous materials), though other light nuclei are also effective. On the *average*, about half the initial neutron kinetic energy is transferred to the protons, so that fast-neutron bombardment looks (to a hydrogenous material) like bombardment with fast protons of half the neutron energy. In addition to such simple collision processes, fast neutrons also induce nuclear reactions in certain elements, leading to emission of particles or photons with their previously described interactions.

After about 20 collisions, neutrons are no longer sufficiently energetic to eject recoil protons but have become "thermalized," i.e., they act (for a short time) like a gas in thermal equilibrium with its surroundings (energies of about 1/40 eV).

When a neutron has become thermalized and wanders into a nucleus, it is quite often captured, momentarily yielding an excited isotope of the original nucleus. Nuclei usually lose their excitation by emission of particles or characteristic gamma rays. Thus, even an uncharged slow neutron gives rise to the release of ionizing radiation inside a material being irradiated. In living tissue, slow neutrons commonly are captured by H^1 nuclei with emission of an energetic gamma ray, and by N^{14} nuclei with emission of an energetic proton.

Associated or Post-ionization Events. Subsequent to the primary and secondary molecular ionizations (and excitations), a number of events can occur that depend rather strongly on the form of the target material, including

(1) dissociation of molecules and formation of free radicals (species with unpaired electrons, hence great chemical reactivity, e.g., in water, H , OH , HO_2);

(2) recombination of ions and radicals, leading to no net change;

(3) dispersion of energetic ions and radicals, and reaction with other species present or with each other;

(4) nondiffusion migration of electronic excitation to energy "sinks," e.g., in macromolecules or crystals; and

(5) eventual degradation of the excess ab-

sorbed energy to heat (insignificant from the standpoint of effects).

For many systems the basic interactions must be considered only the initiators of a complex sequence of later events.

Comparison of the Radiations. Diverse electromagnetic and particulate radiations thus have in common as their basic interaction (or closely following upon it) the production of molecular ionizations (and excitations, dissociations, free radicals) inside target matter. They differ in the *geometry* of these events (especially in LET), a difference that leads to wide variations in range or penetrating ability and, thus, in the subsequent reactions leading to the final expression of the radiation effect.

The slower, more highly charged, heavy particles (such as alphas) travel in straight-line tracks ionizing densely along the track and exhibiting discrete ranges characteristic of the particle energy. Lighter, less highly charged particles such as electrons have their tracks more easily deflected and therefore have less precisely specified ranges in matter (although they have a maximum range). With a lower ionization density, however, they travel much farther than heavy charged particles of the same energy. Gamma or x-rays interact causing release of electrons in matter but at widely spaced intervals and in a random fashion; they have a still lower ionization density (LET) and much longer "range." (Since the resultant of all their absorption processes is a roughly exponential attenuation in matter, their "range" must be described in terms of a parameter such as "half thickness," x_1 , the thickness of material that reduces incident intensity to 50 per cent. The relation $x_1 = 0.693/\mu$ is seen to follow from the above equation for total photon absorption.) X-ray interaction has been likened to a "shotgun" effect in contrast to the "rifle" effect from incident heavy charged particles. Neutrons, which only interact with nuclei, may have either high or low LET. Their penetration in matter is great though difficult to specify well except in terms of specific materials. For much of their path, *fast* neutrons also are attenuated roughly exponentially.

The approximate range r or half thickness x_1 is given in Table 2 for several 1-MeV radiations in water.

TABLE 2

$\alpha(r)$ (cm)	$\beta-(\max r)$ (cm)	$\gamma(x_1)$ (cm)	Neutrons (x_1) (cm)
0.0007	0.4	10	5-10

Finally, a variety of conventional particle and wave interactions (such as reflection, transmission, and refraction) are also experienced by the above radiations. Moreover, many other more-or-less-ionizing radiations have been omitted from this discussion, including mesons, hyperons, heavy cosmic particles, large fission and

spallation products, and anti-particles. Their relative interactions can be quite well predicted from their composition, charge, and velocity, and the above considerations.

HOWARD C. MEL
PAUL W. TODD

Cross-references: ATOMIC PHYSICS, BREMSSTRAHLUNG, COMPTON EFFECT, ELECTRON, IONIZATION, NEUTRON, NUCLEAR REACTIONS, NUCLEAR STRUCTURE, PHOTO-ELECTRICITY, PHOTON, POSITRON, PROTON, RADIATION CHEMISTRY, RADIOACTIVITY, X-RAYS.

RADIATION, THERMAL

The term thermal radiation refers to the electromagnetic energy that all substances radiate, by transformation of their thermal energy. If a body lacks a source of replenishment for the thermal energy thus transformed, it will radiate away all of its available energy, and its temperature will approach absolute zero. As long as the temperature remains above approximately 500°C, some of the thermal radiation will lie in the visible spectrum. At lower temperatures, the energy lies at wavelengths too long to be seen.

The discovery that radiation existed outside the visible spectral region was made by Sir William Herschel in 1800. He formed a prismatic solar spectrum on a table top in a dark room and found that the temperature, as indicated by a sensitive thermometer, continued to increase beyond the red end of the spectrum.

Sources other than the sun were also studied. One that is of fundamental importance is an isothermal enclosure, or hohlraum, with a small hole for viewing the radiation escaping from the interior. Kirchhoff had proved, from the second law of thermodynamics, that the flux and the spectral distribution of the radiation are the same in all such enclosures at a given temperature, irrespective of the materials composing them. A related fact, known as Kirchhoff's law, is that the ratio of radiant emittance to absorptance is the same for all surfaces at the same temperature. (The radiant emittance of a surface is the integrated power radiated in all directions per unit area of surface. Definitions of other radiometric quantities, symbols and units will be found in the references.) In accordance with this law, metallic surfaces, having high reflectance and hence low absorptance, also have a low radiant emittance. Hence, the inner surfaces of the double walls of a Dewar flask are silvered, to minimize radiant heat loss.

At a given temperature, the surface having maximum absorptance will also have maximum emittance of radiant energy. Since such a surface absorbs all incident energy, it appears black, and the radiation it emits is known as "black-body radiation". Like the pupil of the eye, an opening in the surface of a hollow body appears to be perfectly black (as long as there is no reflection back out) and the opening acts as a black-body radiator as well as absorber. No surface

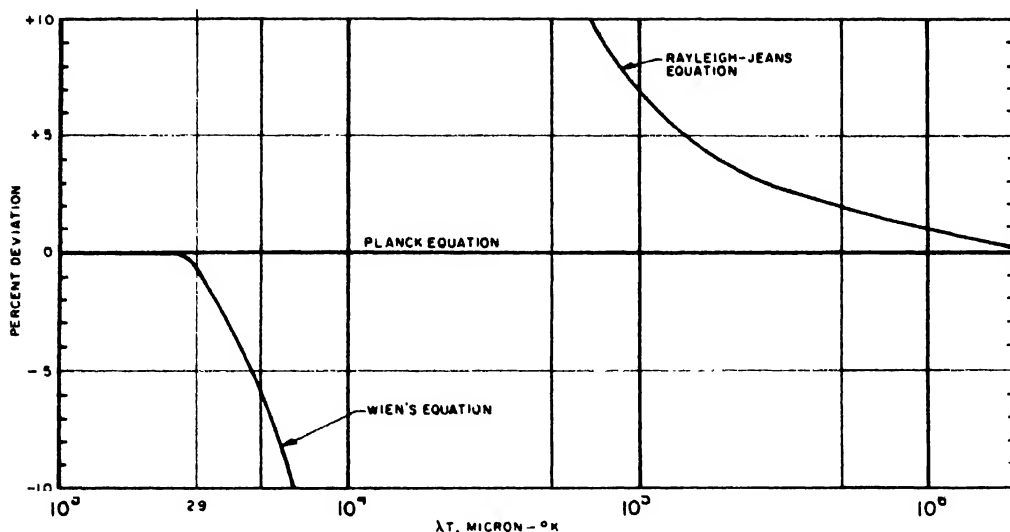


FIG. 1. Fractional deviation of classical radiation equations from the Planck equation.

can have a larger radiant emittance than a blackbody, at a given temperature.

Toward the end of the nineteenth century, quantitative studies of the magnitude and spectral distribution of blackbody radiation were made (see INFRARED RADIATION for typical curves). Stefan had found experimentally in 1879 that the radiant emittance of a blackbody, integrated over all wavelengths, is proportional to the fourth power of the absolute temperature. And, in 1884, Boltzmann gave a theoretical derivation for what is now known as the Stefan-Boltzmann law: $W = \sigma T^4$. The value of the constant σ is 5.67×10^{-12} watts $\text{cm}^{-2} \text{deg}^{-4}$. In 1893, Wien derived a "displacement law," one of whose implications is that $\lambda_{\text{max}} T = 2898$, where λ_{max} is the wavelength (in microns) at which maximum radiance occurs for a blackbody at absolute temperature T .

In 1896, Wien derived the following distribution law for the spectral radiant emittance W_λ of a blackbody: $W_\lambda = c_1 \lambda^{-5} \exp(-c_2/\lambda T)$. This fits the experimental observations within 1 per cent, provided λT is less than 3100, i.e., $\lambda \leq \lambda_{\text{max}}$; but at larger values, the predicted values rapidly become too low (see Fig. 1). On the other hand, by applying the classical equipartition theorem of statistical mechanics to the radiation, Rayleigh and Jeans derived the following formula: $W_\lambda = (c_1/c_2)\lambda^{-4}T$. For $\lambda \geq 250\lambda_{\text{max}}$, the Rayleigh-Jeans equation matches the experimental data to 1 per cent, but it diverges, leading to the "ultraviolet catastrophe," as $\lambda \rightarrow 0$. It remained for Max Planck in 1900 to find an expression that is valid at all wavelengths and temperatures, namely: $W_\lambda = c_1 \lambda^{-5} / [\exp(c_2/\lambda T) - 1]$. c_1 and c_2 are known as the first and second radiation constants, respectively, and have the values $c_1 = 3.74 \times 10^{-12}$ watt cm^2 ; $c_2 = 1.44$ cm degree. Planck's theory leads to the following expressions for the radiation constants in terms of fundamental

physical quantities: $c_1 = 2\pi hc^2$; $c_2 = hc/k$, c being the velocity of light, and k Boltzmann's constant. The validity of these expressions has been experimentally justified.

Planck first developed this law empirically, and tried unsuccessfully to justify it on the basis of classical physics. He was forced to postulate that the elementary oscillators, of which a radiating body consists, did not have a continuous distribution of energy, but only quantized values. According to the "quantum hypothesis," the energy E that an oscillator may assume is given by $E = nh\nu$, where n is an integer, and the proportionality constant h (now known as Planck's constant) $= 6.62 \times 10^{-34}$ watt sec². Though forced to assume that the energies of the elementary oscillators, of which the radiator is composed, could take on only discrete values, Planck considered the emitted radiation to be propagated according to classical electromagnetic theory. The quantization of the radiant energy into photons was conceived by Einstein and used by him to explain the phenomena of photoelectricity. These quantum concepts were later extended by Bohr and Sommerfeld to explain atomic spectra and by Schrödinger, Heisenberg, Dirac, and others to develop quantum mechanics. Thus, they stand as one of the most important milestones in the history of theoretical physics.

Integration of Planck's equation leads to the Stefan-Boltzmann equation. Wien's laws and the Rayleigh-Jeans law may also be derived from Planck's law under appropriate conditions. For example, when dealing with photons whose energy $h\nu$ is sufficiently small compared to the thermal energy kT , the product λT is large enough so that the exponential in the denominator of Planck's law may be replaced by the first two terms of a series expansion. This leads to the Rayleigh-Jeans law.

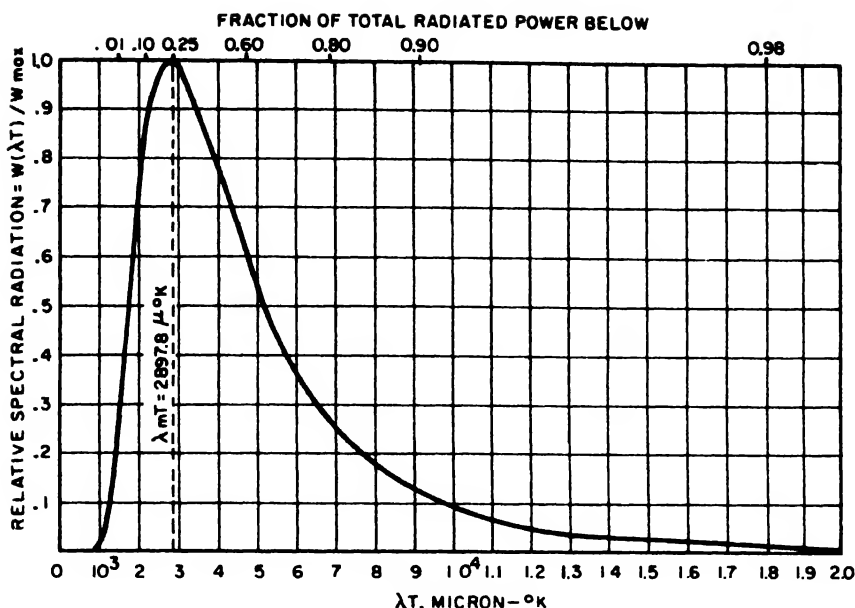


FIG. 2. General blackbody radiation curve.

High-speed computers have been used to compile tables of spectral radiance and related functions for a blackbody over a broad range of wavelengths and temperatures. There are also several radiation calculators, or slide rules, that are convenient for engineering use. Some of these make use of the fact that a log-log plot of blackbody radiation vs wavelength has the same shape at all temperatures. The radiance curve for a given temperature blackbody may then be obtained by sliding the universal curve in such a way that its peak falls on the line corresponding to the Wien displacement law (see Fig. 2 in INFRARED RADIATION). It is convenient to remember that the peak radiance varies as T^5 ; that a quarter of the total power radiated lies between $\lambda = 0$ and $\lambda = \lambda_{\max}$ (see Fig. 2); and that this power, on the short-wavelength side of the peak of the radiation curve varies as $T^{6.4}$. At wavelengths less than λ_{\max} , the monochromatic radiance changes as T^n , where $n \approx 15000/\lambda T$. For an object at 1500°K, for example, observed through an optical pyrometer with a filter transmitting at 0.6μ , $n \approx 17$. This relatively high value of n is advantageous in reducing the temperature error resulting from uncertainty in the observed radiance or the emittance of a surface.

The ratio by which the surface radiation falls short of that of an ideal blackbody, is denoted as the surface's emissivity or emittance. Some authorities prefer to use "emissivity" as a material property characterizing an ideally pure and polished surface of the material, as distinguished from the properties of engineering samples. The emittance of some substances, though less than one, is essentially the same at all wavelengths. Such radiators are referred to as gray bodies. In most real substances, however, emittance varies

as a function of wavelength as well as temperature.

If one can make a reasonable estimate of the emittance of a surface, a measurement of its thermal emission will give its temperature. This is the basis of contactless, radiometric temperature measurement. Should the emittance be unknown, but assumed to be relatively constant, a temperature may be inferred from the shape of the spectral distribution curve. Other schemes have also been used to determine the temperature of surfaces of undetermined characteristics.

The *color temperature* of a body is that temperature of a blackbody which has the same ratio of radiances as the selective radiator in two spectral intervals. In general, the value of the color temperature depends on the choice of the two spectral intervals. However, if these wavelengths are in the visible region, the color temperature is relatively insensitive to the specific values chosen. The color temperature of a gray body equals its true temperature. When dealing with semitransparent materials such as glass in the near infrared, one speaks of volume emissivity, which in turn is related to the optical constants of the material.

In some solids, and especially in gases and flames, the emittance in some spectral regions is much larger than in others. Kirchhoff's law tells us that gases radiate well in the same spectral regions where they have strong absorption bands. One must bear in mind, however, that the shape of an absorption band changes and new "hot" bands may appear as the gas temperature rises, so emittance and absorptance must be equated at the same temperature. The thermal radiation from a cool, low-pressure gas, is resolvable into discrete emission lines, as explained by the

quantum theory. Increased temperature and pressure cause the lines to broaden as a result of molecular interactions and perturbations of the energy levels, and new lines also appear. As the optical path length increases, the emissivity of the gas approaches unity, at first at the line centers, where the absorptance is strongest, and then extending into the line wings. Eventually, when the gas density and thickness are great enough, as on the sun, the original line spectrum assumes the appearance of a blackbody distribution. A good example of this is seen when one compares two infrared emission spectra of earth's atmosphere—first looking overhead from a mountain top, and then along the horizon at a humid sea-coast. The latter spectrum's blackbody-like shape (from which, incidentally, the atmosphere's temperature may be inferred) contrasts strongly with the peaks and valleys of the former.

In strongly absorbing spectral regions, only a relatively thin layer of gas, nearest the viewer, contributes to the observed radiance, the radiation by the more distant molecules having been almost completely absorbed by the nearer molecules. For this reason, the Fraunhofer lines in the solar spectrum, occurring at strongly absorbing atomic wavelengths, originate in the cooler outer part of the sun and appear relatively dark.

In exceptional situations, it is possible to circumvent the consequences of Kirchhoff's law. The Doppler shifts, due to the relative motion of the earth and Mars, prevents the narrow line radiation by planetary H_2O from being reabsorbed by terrestrial water vapor. As a result, it has been possible to estimate the water vapor concentration on Mars (see DOPPLER EFFECT AND PLANETARY ATMOSPHERES).

The spectral emittance of gases is an important area of investigation, with respect to such topics as the heat budget of the earth; the composition of planetary atmospheres; and radiation by flames and rocket engines. The analysis of the gas radiation transfer in many of these applications is complicated by the fact that conditions are nonisothermal.

The Welsbach mantle (or the Coleman lantern) is an example of selective radiation in a solid; in this case, its efficiency as a source of visible light is enhanced by the high emittance in the visible, and low emittance in the near infrared, of the mixture of thorium and cerium oxide of which the source is composed. For some metals, the emittance varies as $\sqrt{T/\lambda}$, while in others it may vary in a more complicated way. In either case, the form of the spectral distribution curve differs from that of a blackbody, and the total emission will usually vary more nearly as T^5 than as T^4 . The emittance of a tungsten filament is approximately 0.45 in the visible spectrum, decreasing to less than 0.2 in the infrared. Its visible efficiency increases as its operating temperature increases.

By considering the equilibrium between incoming and outgoing radiation, one can show that the sum of the radiant absorptance a , reflectance r , and transmittance t , is unity. These quantities refer to monochromatic, hemispherical radiation

from the surface and do not take into account the way it is distributed geometrically. For an opaque surface, $a + r = 1$, and since by Kirchhoff's law $a = e$, the emittance, we have $e = 1 - r$. Errors may result if due account is not taken of angular and spectral factors in the application of this equation. Early studies of the angular distribution of radiation from surfaces led to the formulation of Lambert's law, i.e., $J_\theta = J_n \cos \theta$, where J_θ is the source intensity or radiant power per unit solid angle in a direction making an angle θ with the normal to the surface, and J_n is the intensity along the normal. A surface that obeys Lambert's law is said to be perfectly diffuse, like a sheet of blotting paper in the visible. Its radiance, N (intensity per unit of projected area of source), is constant and independent of θ . Integration over the hemisphere leads to the relationship $W = \pi N$. A truly black surface obeys Lambert's law exactly. Although the law is a useful approximation for many radiators and reflectors, there are numerous exceptions. The emissivity of clean, smooth surfaces of some materials has been studied from a basic theoretical viewpoint and related to their optical constants. The analysis, whose results agree generally with experimental determinations, indicates that for electrical conductors, the normal emissivity is quite low and increases with θ ; whereas insulators have relatively high normal emissivity, decreasing at large values of θ . Insulators, as well as metals covered with thick oxide films, behave approximately as diffuse radiators. In the case of many practical materials, however, the surfaces are either rough or chemically complex, and it is necessary to rely upon empirically determined emittances. Many substances, such as concrete, porcelain, and paper, have a higher absorptance for long- than for short-wave radiation. Many more data are needed on the spectral emittances of various materials under different conditions. In addition to their engineering uses in radiant heat transfer calculations, emittance and reflectance data are being used to deduce the chemical composition of the moon and planets.

Space vehicles absorb energy from the sun and lose it by radiation to space. By the application of coatings with suitable radiative characteristics, the internal temperature of a vehicle may be controlled with desired limits. The quantity a/e is often used to characterize such coatings. It refers to the ratio of the absorptance of solar radiation (approximated by a 6000°K blackbody) to the emittance at the temperature of the vehicle's surface. High-temperature emittances of many materials are being studied in connection with the design of ablative nose cones, and re-entry vehicles.

Although all objects are continuously emitting and absorbing thermal radiation, specially designed sources are available for particular applications. For industrial heating and drying, for example, there are numerous varieties of heaters and infrared lamps. The latter are similar to incandescent lamps used for lighting purposes, but designed for operation at lower temperatures,

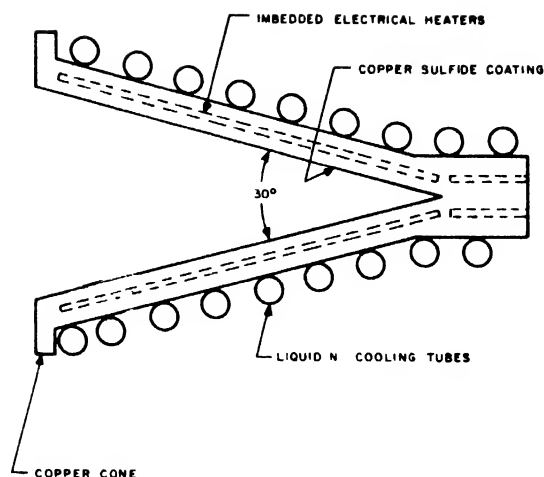


FIG. 3. Schematic representation of a conical blackbody that can be used at high or low temperature.

with reduced visible output. In recent years, high-temperature lamps with quartz envelopes have been used increasingly for both heating and illumination. The clarity of the envelope is maintained, and the filament life prolonged, by incorporating a small amount of iodine within the tube. Tungsten that has evaporated onto the tube walls combines with the iodine vapor to form tungsten iodide. This is a gas which decomposes thermally when it comes in contact with the hot filament, thus replenishing the latter and liberating the iodine for use again.

For scientific purposes, such as calibrating a radiometer or spectrometer, a reproducible source of known characteristics is essential. Sources with characteristics approaching those of an ideal blackbody, have been built for operation from cryogenic temperatures to about 3000°K. Most of the commercially available blackbodies, which come with a variety of aperture sizes and in many configurations, use a conical cavity (Fig. 3), the interior walls of which are oxidized or blackened to decrease their reflectance. Electrical heating is most often used, in conjunction with a thermostatic controller. A number of authors have discussed the considerations involved in designing a blackbody, and it is possible to calculate how closely a given design approximates an ideal blackbody. It is sometimes necessary to have a black radiating surface whose area is impractical to obtain with a cavity-type body. Flat surfaces have been coated with "blacks" whose emittances are close to unity over a broad spectral range. These are generally not suitable for high-temperature use, however. Other methods that have been used to obtain high-emittance surfaces include: wedges or closely stacked razor blades viewed edge on, a telescoped cone pattern similar to that of a Fresnel lens impressed on a flat surface, a vortex

in a liquid or molten metal, etc. The National Bureau of Standards has calibrated some special tungsten filament lamps for use as spectral irradiance standards from about 0.3 to 2.6μ . Beyond 2.6μ , strong, but variable, atmospheric absorption becomes troublesome. When using infrared sources for quantitative thermal radiation measurements, one must make due allowance for the absorption at different infrared wavelengths by the CO_2 and water vapor in the atmosphere between the source and the instrument being calibrated.

For reasons of compactness and convenience, it is sometimes desirable to use non-blackbody sources of thermal radiation. Among the most common of these are the Nernst glower, which is a hollow rod approximately 25 mm long and 2 mm in diameter, made of a mixture of oxides of zirconium, yttrium, and thorium and the globar, a rod of bonded silicon carbide. These are most useful at the shorter infrared wavelengths. In the far infrared, mercury discharge tubes and other sources have been used. At long wavelengths, the power output of a thermal radiator increases almost linearly with temperature, as indicated by the Rayleigh-Jeans law, and inordinately high temperatures would be required for a significant increase in power. Nonthermal sources, in particular lasers, or other coherent radiators can generate much greater power, though limited to a narrow wavelength interval.

In order to study the properties of materials at very high temperatures without contamination, it is convenient to heat them by thermal radiation rather than convection. Solar furnaces have been used for this purpose, as well as arc-imaging furnaces, in which a specially designed high-intensity carbon arc replaces the sun as the source.

Thermal radiation can be detected by eye when the source temperature is sufficiently high. And

although its sensitivity is greatly reduced at longer wavelengths, the human eye has some sensitivity to radiation at wavelengths as long as 1.2μ . Some insects, such as moths, respond to thermal radiation at even longer wavelengths, and some snakes can detect the heat of a warm-blooded animal at an appreciable distance.

Many types of detectors have been developed for thermal radiation. In some of these, known as quantum detectors, the energy of the incident photon must exceed some lower limit, but its effects are very rapid. An example is the lead sulfide detector. In others, thermal radiation of any wavelength can be detected by virtue of its effect on some physical property of the detector. The response of the latter kind of detector is slower than that of the former. Among the thermal detectors are: metal and thermistor bolometers, in which the absorbed radiation causes a slight temperature rise and consequent change in the resistance of the detector; thermocouples and thermopiles in which the differential heating of two junctions of dissimilar materials generates an emf; the Golay cell or pneumatic detector, where the absorbed radiation heats a gas and distorts a reflecting optical element; the evaporograph in which the incident radiation causes differences in the rate of evaporation of a thin oil film and the resulting differences in thickness give rise to interferometric patterns. Numerous other physical mechanisms have been exploited for radiation detection. (The more usual infrared detectors are discussed under INFRARED RADIATION.)

For coverage of an extended field of view, mosaics of conventional detectors have been used, as well as electronic imaging tubes such as infrared vidicons and orthicons. Whether by such detectors or by raster-like scanning of a small elemental field of view, it is possible to study the temperature pattern of an extended source of thermal radiation. Thermography is such a process, in which the temperature gradients are displayed in visible form, e.g., on photographic film. When properly interpreted, the thermograph can be a valuable diagnostic aid to the physician. Industrial thermography, as applied, for example, to the nondestructive testing of integrated electronic circuit boards is useful for design purposes or for quality assurance. In interpreting the observed patterns, one must bear in mind that radiance variations may result from differences in either or both surface emittance and surface temperature from point to point. This fact is less troublesome in medical thermography because the emittance of the body is close to unity. In a similar way, the radiometry of lakes, rivers, and oceans by an instrument looking vertically down from an aircraft, takes advantage of the near-unity emissivity of water within the range of wavelengths to which the radiometer is sensitive. In the near infrared and visible, as well as in the microwave region, water is not quite so opaque as it is around 10μ . Microwave radiometry, therefore, can give information about the water temperature slightly (1 or 2 mm) below the surface, rather than at the surface itself. Microwaves

are less subject than infrared to atmospheric absorption and are better able to penetrate overcast and clouds.

Satellite observations of the thermal radiation of the oceans and continents are of meteorological value. A special radiometric spectrometer, currently under development, records the thermal radiation of atmospheric carbon dioxide in a series of narrow wavelength intervals centered around the strong 15μ absorption band. The varying emittances of the atmosphere in these wavelength regions permit one to deduce a vertical profile of the atmosphere's properties from the observed data.

LEONARD EISNER

References

- Forsythe, W. E., "Measurement of Radiant Energy," New York, McGraw-Hill Book Co., 1937.
- Worthing, A. G., and Halliday, D., "Heat," New York, John Wiley & Sons, 1948.
- Richtmyer, F. K., and Kennard, E. H., "Introduction to Modern Physics," Fifth edition, New York, McGraw-Hill Book Co., 1955.
- Klein, M. J., "Max Planck and the Beginnings of the Quantum Theory," in *Archive for History of Exact Sciences*, 1 (5), 459-479 (1962).
- Pivovonsky, M., and Nagel, M. R., "Tables of Black-body Radiation Functions," New York, The Macmillan Co., 1961.
- Rutgers, G. A. W., "Temperature Radiation of Solids," Flugge, S., Ed., *Encyclopedia of Physics*, Vol. 26, pp. 129-170, Berlin, Springer, 1958.
- Harrison, W. N., et al., National Bureau of Standards, "Standardization of Thermal Emittance Measurements," WADC Technical Report 59-510, in 4 parts 1960-64, Office of Technical Services, U.S. Dept. of Commerce.
- Gubareff, G. G., Janssen, J. E., and Torborg, R. H., "Thermal Radiation Properties Survey," Second edition, Minneapolis-Honeywell Regulator Co., Honeywell Research Center, Minneapolis, Minn., 1960.
- "Measurement of Thermal Radiation Properties of Solids," A Symposium held at Dayton, Ohio, Sept. 5-7, 1962, NASA SP-31, for sale by U.S. Govt. Printing Office.
- "Symposium on Thermal Radiation of Solids," University of California, San Francisco, California, March 4, 5, 6, 1964.
- Penner, S. S., "Quantitative Molecular Spectroscopy and Gas Emissivities," Reading, Mass., Addison-Wesley Publishing Co., 1959.
- Sobolev, V. V., "A Treatise on Radiative Transfer," Princeton, N.J., D. Van Nostrand Co., 1963; translated by S. F. Gaposchkin.
- Gebhart, B., "Heat Transfer," New York, McGraw-Hill Book Co., 1961.
- Glaser, P. E., and Walker, R. F., Eds., "Thermal Imaging Techniques," New York, Plenum Press, 1964.
- Baker, H. D., Ryder, E. A., and Baker, N. H., "Temperature Measurement in Engineering," 2 vols., New York, John Wiley & Sons, 1953 and 1961.

American Institute of Physics, "Temperature, Its Measurement and Control in Science and Industry," Vols. 1-3, New York, Reinhold Publishing Corp., 1941-1963.

Goody, R. M., "Atmospheric Radiation. I. Theoretical Basis," Oxford, Oxford University Press, 1964.

Summer, W., "Ultraviolet and Infrared Engineering," New York, Interscience Publishers, 1962.

Cross-references: DOPPLER EFFECT; HEAT TRANSFER; INFRARED RADIATION; KINETIC THEORY; LIGHT; OPTICS, GEOMETRICAL; OPTICS, PHYSICAL; PHOTON; PLANETARY ATMOSPHERES; QUANTUM THEORY; REFLECTION; SOLAR PHYSICS; SPECTROSCOPY; STATISTICAL MECHANICS; TEMPERATURE AND THERMOMETRY.

RADIO ASTRONOMY

Radio astronomy is concerned with measurement of radio-frequency radiation emitted naturally by celestial objects. It began in 1932, with the discovery of radio emission from the galaxy by K. G. Jansky at the Bell Telephone Laboratories. Fourteen years later, the U.S. Army Signal Corps and the Hungarian scientist Z. Bay, working independently, succeeded in obtaining radar echoes from the moon—thereby launching the related field of *radar astronomy*. Most recently, F. T. Haddock and D. Walsh at the University of Michigan, and A. E. Lilley at Harvard, have pioneered in *space radio astronomy*—the use of rocket-borne or orbiting radio telescopes to detect low-frequency radio waves that cannot penetrate the terrestrial ionosphere.

Radio Telescopes. A radio telescope consists basically of one or more aerials and a radiometer. The chief factor taken into account in the design of these components is the frequency range within which observations are to be made, and this parameter in turn depends on the nature of the studies proposed for the telescope. Investigations at meter wavelengths are chiefly limited by confusion of adjacent radio sources, rather than by instrumental sensitivity. Therefore, the emphasis in designing telescopes for use at these frequencies has been on the attainment of high spatial resolution, and this requirement has generally led to the adoption of interferometers in preference to "single dish" or filled aperture antennas.

L. L. McCready, J. L. Pawsey and R. Payne-Scott at Sydney introduced the technique of interferometry to radio astronomy in 1946. They used the Lloyd's mirror principle, with the ocean surface as reflector, to measure the angular diameter and position of a localized source of solar radio emission.

Two-element interferometers have been employed for sky survey work at the Mullard Radio Astronomy Observatory; others, notably those at Jodrell Bank, Owens Valley and Nançay, have yielded basic data on the brightness distributions of individual sources by means of variable baseline interferometry. As greater resolution becomes necessary to pursue current lines of inquiry at centimeter wavelengths, the two-element inter-

ferometer will become popular at those frequencies as well. Indeed, such an instrument, comprising two precisely figured 85-foot paraboloids, was completed in the summer of 1964 at the National Radio Astronomy Observatory, Green Bank. The simple interferometer has a *fan beam*, i.e., its high resolution is limited to one coordinate. Symmetrical or *pencil beams* have been synthesized at meter wavelengths with two-dimensional arrays of interferometer elements such as the *Mills Cross* developed at Sydney. However, the need for better resolution has already led to the construction of such arrays with characteristic dimensions on the order of a mile. The large number of individual aerials required in these projects raises many problems, not the least of which is cost. *Aperture synthesis*, an alternative approach introduced by M. Ryle and A. Hewish at Cambridge, involves the use of two or more aerials, at least one of which is movable, to simulate a much larger array. Data must be taken with the aerials in each of many different relative orientations, in order to obtain one "observation."

The 250-foot "dish" at Jodrell Bank is the largest fully steerable parabolic antenna. However, its surface is not suitable for use at centimeter wavelengths, and most of the important results in this frequency range have been obtained with smaller dishes, such as the 50-foot reflector at the Naval Research Laboratory and the nearly identical 85-foot reflectors at Michigan and Green Bank. Three large "filled aperture" telescopes have recently been put into operation; two, the 210-foot fully steerable dish at Parkes, N.S.W., and the 300-foot transit telescope at Green Bank, are paraboloids, and the other, a 1000-foot reflector constructed in a natural limestone sinkhole at Arecibo, Puerto Rico, has a spherical surface.

The 72-foot reflector at Serpukhov, U.S.S.R., dominates the field of millimeter wave radio astronomy. Smaller reflectors are in use at these frequencies in California and Texas, and another will be constructed on Kitt Peak, in southern Arizona.

Several approaches to the problem of the mechanical support of large, steerable reflectors have been tried. Among these are the Ohio State University instrument, which consists of a fixed parabolic section and a tiltable flat reflector, the parabolic strip antenna of Pulkovo Observatory, and the cylindrical paraboloid of the University of Illinois. Many other antenna configurations have been used or proposed by radio astronomers in response to the requirements of various observational programs.

Gain stability, phase stability and low receiver noise temperature are the prime desiderata in a radiometer. Compensation for gain variation by switching between the object signal and a reference standard was achieved by R. H. Dicke in 1946. Many refinements of this technique have since been developed. The problem of phase stability is greatest for long-base-line interferometers and multielement arrays; phase switching

and correlation techniques have generally been used in conjunction with these instruments. Traveling wave tubes, masers, parametric amplifiers and tunnel diodes have all been used to provide low noise amplification, particularly in the microwave region where the average intensity of cosmic sources is low. Multichannel receivers have received wide attention, especially for use in spectral line studies.

Cosmic Radio Sources. In the early stages of radio astronomy, any unidentified celestial radio source was generally called a *radio star*. However, as increased accuracy in the radio positions was attained, more and more radio sources were identified with optical objects, and it became clear that these sources were external galaxies, supernova remnants, etc., not stars. In fact, ordinary stars are very weak radio emitters, and therefore the only member of this class that has been studied by radio astronomers is the sun.

Solar radio waves are classified in terms of three basic components. The *quiet sun* radiation consists of the background thermal emission of the solar atmosphere. The *slowly-varying* component, characterized by a 27-day recurrence tendency, is produced in regions of enhanced density that form in the corona above solar active regions; V. V. Zheleznyakov, and also T. Kakinuma and G. Swarup have shown that it can be explained in terms of radiation by thermal electrons at the gyrofrequency and its harmonics (*resonance absorption*). Coronal temperatures inferred from the quiet sun and slowly-varying components are about 10^6 °K. The *solar bursts* constitute the third component. They may occur as isolated events with durations as short as 10^{-2} second, in groups, or in "storms" that may persist for several days. The bursts that occur in different regions of the radio spectrum differ in their observed properties, as well as in their places and modes of origin. The mechanisms by which the bursts are generated have not been satisfactorily discussed for all types of events; however, the explanation of certain kinds of bursts in terms of plasma oscillations, and others in terms of the synchrotron process, is generally accepted.

It has been known for some time that certain cool ($T \approx 2900$ °K, typically) dwarf stars show large, rapid, nonpredictable increases in optical brightness. Recently, A. C. B. Lovell, F. L. Whipple and L. H. Solomon have studied several of these *flare stars*. They found that radio bursts of much greater absolute intensity than the average solar event are associated with the optical flares of these stars. Analysis of the data revealed that the velocities of light and radio waves are the same to 4 parts in 10^7 ; thus a combination of optical and radio astronomy observations has provided physicists with what is at present the most accurate experimental proof for the non-wavelength-dependence of the velocity of electromagnetic radiation.

During the past decade, interest in planetary astronomy was revived, first by the results of radio observations and subsequently by the advent of the "Space Age." The first radio detection of

Mercury was made in 1961 by W. E. Howard, A. H. Barrett and F. T. Haddock, who used the 85-foot Michigan reflector. The preliminary analysis of observations made with the 210-foot Parkes telescope, recently reported by K. I. Kellermann, suggests that the temperature differential between the light and dark sides of Mercury is small compared to the average value of about 300°K. The simplest explanation for such a result, namely heat transport by atmospheric processes, is in direct contradiction to simple physical theory, which suggests that Mercury should be unable to retain an atmosphere. Extensive observations by many workers have shown that the microwave spectrum of Venus resembles that of a blackbody at the unexpectedly high temperature of 600 K. Observations of limb darkening, made with the Mariner II space vehicle at a wavelength of 1.9 cm, appear to have ruled out the possibility that the microwave emission arises in a dense ionosphere, and at present, a strong *greenhouse effect* seems to be the most reasonable explanation. American, British and Soviet radar astronomers have independently found that Venus rotates in the retrograde sense. As early as 1956, C. H. Mayer, T. P. McCullough and R. M. Sloanaker at the Naval Research Laboratory showed that the brightness temperature of Mars at wavelength 3.15 cm is about 200°K, in reasonable agreement with infrared measurements. In 1964, a temperature of 1100°K at 21 cm was found at Jodrell Bank, but this result has not been substantiated by subsequent work. In 1955, radio bursts from Jupiter were accidentally discovered by B. F. Burke and K. L. Franklin of the Carnegie Institution. Their report led C. A. Shain to identify over 50 previously unidentified Jovian bursts in pre-discovery Australian radio observations; he also inferred the existence of a localized burst source on the planet from the fact the frequency of burst occurrence depended on the longitude of the Jovian central meridian. At present, due chiefly to long series of observations made at Yale, Florida and Colorado, it is known that the Jovian bursts occur in the frequency range below 45 Mc/sec and that they originate in several distinct sources with angular diameters of 15 seconds of arc or less. The occurrence rate of the bursts appears to be correlated with the position of some Jovian satellites. Continuous radio emission, in both the spectral and temporal senses, is observed in the decimeter region. It has a nonthermal spectrum and originates in a region whose equatorial radius is about three planetary radii and whose polar radius is about one planetary radius. Recently, J. R. Dickel at Michigan has extended the observations of this component to a wavelength of 1.8 cm, where thermal emission may also be important. The theoretical explanation for the various components of the Jovian radiation is still a matter of considerable controversy. Several groups have detected Saturn in the wavelength range 3 to 21 cm, but this work is still in a preliminary state. Radio detection of the three outermost planets has not yet been accomplished.

In addition to the well-known radar echo experiments, radio observations of the moon in the wavelength range 4 mm to 75 cm have been carried out extensively since 1946. The radiation is thermal in origin and has generally been studied from three points of view: the comparison of radio emission from different features of the lunar topography, the dependence of the emission on the phase of the lunar cycle, and the study of variations in the emission that occur during eclipses of the moon. The amplitude and phase lag of the emission shows that it originates in a thin region just below the lunar surface, and attempts have been made to interpret the results in terms of layers of dust and rock.

At a 1944 Leiden Observatory colloquium, H. C. van de Hulst predicted that a spectral line could be observed by radio astronomers, who previously had worked only in the continuum. He referred to the 1420-Mc/sec hyperfine transition in the ground state of neutral hydrogen, the now-famous 21-cm line. Seven years later, H. I. Ewen and E. M. Purcell at Harvard made the first successful detection of this radiation. The importance of the 21-cm line is twofold: (1) it originates in the most abundant known constituent of interstellar matter; (2) Doppler shifts are readily measurable in line radiation but not in the continuum. As a result of many 21-cm investigations, notably those of the Harvard, Leiden and Australian groups, it has been possible to map the distribution of interstellar hydrogen in the galaxy and to study the dynamics of the gas. In 1963, A. H. Barrett, M. L. Meeks, and S. Weinreb at M.I.T. succeeded in detecting the 18-cm lines of OH. Already, further studies have shown that the properties of the interstellar gas as determined from the hydroxyl radical differ in several respects from those revealed by the hydrogen line. Hydrogen-line measurements have now also been reported for several dozen external galaxies. Interstellar clouds in the vicinity of hot stars may become photoionized (*H II regions*). Observations of the thermal radiation from such objects, generally made on decimeter and centimeter waves, lead to estimates of temperatures, electron densities, and total masses; further, the structural details revealed by high-resolution observations, such as those made by Y. Terzian at Green Bank, are of great interest to the theoretician. Radio detection of several *planetary nebulas* has also been achieved.

A number of radio sources have been identified with the remnants of galactic supernovae, some of which were unknown prior to the radio observations. Their brightness distributions tend to be circularly symmetric, sometimes with indications of shell structure, but condensations and filamentary detail are also found. The observations are in accord with the suggestion of I. S. Shklovsky that the radiation is produced by the synchrotron process. Measurements of Cassiopeia A, the remnant of a Type II supernova that exploded in about 1700 A.D., have confirmed Shklovsky's prediction that the radio flux should decrease by about 2 percent per year, independent of frequency.

To the radio astronomer, there are two categories of galaxy. A weak emitter, or *ordinary galaxy*, (such as our own system) typically radiates 10^{38} erg/sec in radio waves, compared with 10^{44} erg/sec in the optical region. The strong emitters, or *radio galaxies*, each produce up to 10^{45} erg/sec in the radio range alone. In both types of galaxy, the radio continuum is accounted for on the synchrotron theory. Radio data on ordinary galaxies have been rather sketchy, due to the small number of positive detections. Recently, however, D. S. Heeschen and C. M. Wade at Green Bank have made a systematic study of bright galaxies; they conclude that all normal spiral and irregular galaxies are probably weak radio sources. On the other hand, the strong emitters tend to be elliptical galaxies, often distinguished by peculiar optical phenomena, and they are believed to constitute the bulk of the nearly 1300 discrete sources (most are not yet identified) listed in the General Catalogue of W. E. Howard and S. P. Maran. The early results indicate that these sources have power-law spectra, but the more accurate data now available show the presence of curvature in many spectra (as displayed in a log frequency-log flux diagram, where a power law is a straight line), and W. A. Dent and F. T. Haddock at Michigan have demonstrated the value of observations below wavelength 4 cm in defining the spectral shape. The interferometric studies of A. T. Moffet and P. Maltby at California Institute of Technology and of J. Lequeux at Nançay, enabled them to divide the resolved sources into three groups, on the basis of brightness distribution: *simple*, *double* and *core-halo* sources. C. Hazard and M. B. Mackey at Sydney have used the method of lunar occultations to resolve detail as small as 0.5 second of arc and have detected triple and even more complex structure in some sources. Moffet has recently reexamined several double sources with improved resolution and has found that the emission in a given lobe tends to be concentrated at the end furthest from the other lobe, in accord with an expanding model. Linear polarization measurements at several frequencies have made it possible to study depolarization and the rotation of the electric vector as a function of frequency, leading to model-dependent estimates of the electron density and magnetic field strength in interstellar and intergalactic space as well as in the sources themselves. Similar estimates for the sources are obtained from the radio spectra and angular diameters, combined with distances obtained from application of the Hubble law to optically observed redshifts, according to the method of G. R. Burbidge (see COSMOLOGY).

The identification in 1963 of a new class of radio source the *quasi-stellar objects*, now called "quasars," had a profound impact on physics and astronomy. Originally thought to be peculiar stars within our galaxy, the quasars are now known to be among the most distant objects in the universe. Although their radio emission typically is similar to that of a strong radio

galaxy and the optical emission rates are up to 100 times greater than those of any known galaxy, they occupy volumes that are extremely small compared to typical galactic dimensions. Detailed optical studies of quasars are mostly limited to the very largest telescopes. A. R. Sandage at Mount Wilson and Palomar Observatories has recently discovered several dozen of these objects, but our knowledge of them is limited, and the theoretical explanations of the vast energies involved are still in the speculative stage.

Radio Astronomy and Cosmology. The several methods that are suitable, in principle, for testing between world models on the basis of radio observations include (1) the measurement of background extragalactic radiation, (2) the determination of the number of sources in each flux-density range, and (3) the analysis of the relation between the angular diameters and the flux densities of sources. However, these tasks are very difficult, and the only reasonably certain result thus far obtained is that of M. Ryle, who found that the number-flux density relation is not in accord with the basic version of the steady-state theory.

STEPHEN P. MARAN

References

- Maran, S. P., and Cameron, A. G. W., Eds., "Physics of Nonthermal Radio Sources," N.A.S.A. SP-46, U.S. Government Printing Office, 1964.
Steinberg, J. L., and Lequeux, J., "Radio Astronomy," New York, McGraw-Hill Book Co., 1963.

Cross-references: ANTENNAS, ASTROPHYSICS, COSMOLOGY, INTERFERENCE AND INTERFEROMETRY, SOLAR PHYSICS.

RADIOACTIVE TRACERS

The use of radioisotopes as tracers rests on the nearly indistinguishable physical and chemical properties of all the isotopes of a given element. Proper incorporation into a material of a radioisotope, that can be measured with appropriate radiation detectors, provides a means of studying the behavior of the material or a component thereof. Thus, a radioisotope may be used to study the chemistry or physics of an element, a chemical compound, or a mixture of substances. For example, I^{131} , a radioisotope with an eight-day half-life, has been used to study the distribution of iodine in multiple phase systems, to study the biochemistry of I^{131} -tagged diiodotyrosine, and to measure the flow rates of underground streams.

Shortly after the end of World War II, the U.S. Atomic Energy Commission made a variety of radioisotopes available from the Oak Ridge National Laboratory. Subsequently, additional suppliers have been established in the United States and in other countries. One may now purchase useful radioisotopes of 68 out of the 81 "stable" elements. Isotopes of very short half-life

(minutes to an hour or so) exist for another 10 "stable" elements. Useful radiotracers are completely absent only for the elements, He, Li, and B. Most radioisotopes are produced in nuclear reactors by fission or other neutron-induced reactions; however, some are produced with accelerators. Available radioisotopes, their sources and their prices are catalogued in "The Isotope Index, 1963-1964," J. L. Sommerville, Editor, Scientific Equipment Co., Indianapolis, Indiana, 1963.

Facilities for the formation of radioactive tracers directly within a sample or test material have also become available. For example, it is possible to form radioactive iron 59 within a sample of steel by irradiating the sample in a nuclear reactor.

The wide choice of radioisotopic tracers and the availability of sensitive detection systems to fit most circumstances have made possible the use of radiotracer techniques in many branches of science, medicine, and industry. Radiotracers have a number of features that make their use generally attractive. Unlike other types of tracers, they provide unequivocal evidence of their presence by virtue of their own radiation. Most species of radioisotopes are inexpensive (although incorporation into a specific compound can be somewhat costly), and detection equipment can be obtained at moderate expense. Also, due to the excellent detection efficiency of available equipment, it is possible to carry out most experiments without undue health hazards. Investigators trained in the safe handling of radioisotopes can follow gas, liquid, solid, or mixed-state systems at the laboratory, pilot-plant or even full plant scale with complete safety.

The following examples of radiotracer applications comprise but a partial list of uses.

Absorption of gaseous or liquid-phase constituents can be readily studied with the aid of radiotracers. Either the deposition of tracer onto the substrate or the disappearance of tracer from its initial phase may be measured.

Analysis of chemical composition may be accomplished either by utilizing the reaction of tagged reagents or by "isotope dilution." An example of the former method is the use of Ag^{110} -tagged silver nitrate reagent in chloride determinations by the precipitation of $AgCl$ from solution; a sharp rise in liquid-phase radioactivity indicates the point of essentially complete chloride precipitation. Isotope dilution utilizes an isotope of an element to measure the amount of the same element in a sample. It is based on the fact that the amount of radioactivity per unit weight of the element or compound (specific activity) in the tracer reagent will be decreased when the reagent is added to a solution containing the naturally-occurring element or compound. The change in specific activity is an analytical measure of the amount of the element or compound originally present in the sample.

Radiotracers may be used as analytical adjuncts, also. They provide a convenient means of checking the degree of completion of analytical steps, such

as precipitation and extraction, and can provide a measure of chemical losses in analytical procedures. They may be used to mark compound locations in chromatographic separation procedures.

Many other aspects of chemistry have been elucidated with radiotracer techniques. Chemical reaction rates, equilibria, and mechanisms have been studied. Diffusion rates, exchange rates, solubility products, partition coefficients, dissociation constants, and vapor pressures have been measured. Processes due to the effects of high-energy radiation, photolysis, and catalysis have been unraveled. Many of the recent advances in biochemistry would not have been possible without the use of radioactive tracers, especially in the study of biological catalysis (enzyme reactions).

Many of the above subjects have been studied in connection with fields other than chemistry. Thus, while diffusion rates are of interest in elucidating rate-limiting chemical processes, they are also of considerable interest in electronics, metallurgy, and process industries. The self-diffusion of alloy components as a function of alloy composition and grain substructure has been investigated. Diffusion in other solid-state materials, such as semiconductors, has been measured. The specific sulfide surface area of metal sulfides supported on alumina has been determined by exchange of normal surface sulfur with S^{35} -tagged H_2S . The rate of such exchange is of interest, and the final equilibrium state gives a measure of the surface sulfide area. These are examples of problems that are not amenable to solution with techniques other than the radiotracer method.

Industry has obtained marked economic benefits from radiotracer applications. Corrosion and wear studies can be carried out with rapidity and insight otherwise impossible. The corrosion of a steel pipe containing Fe^{59} (produced by irradiating a section of pipe) can be followed *in situ* by measuring the appearance of radioactivity in the corrosive medium or by following the disappearance of radioactivity of the part. Similarly wear of an irradiated part such as a piston ring, cylinder sleeve, or gear, can be measured *in situ*. Prior to the availability of radiotracers, such wear studies required frequent dismantling of machinery for weight measurements. Furthermore, it is now often possible to obtain detailed wear or corrosion patterns by autoradiography. The techniques used in wear studies have also been used to study the effectiveness of lubricants and the mechanism of wear prevention.

Radioactive tracers have been used to examine fluid processes. They are highly useful in detecting leaks and are used routinely to mark the interface between two different products moving consecutively through a pipe-line. The gamma rays from an isotope such as Ba^{140} in soluble form at the interface can be discerned easily through the walls of the pipe. In a similar fashion, flow rates may be measured by quickly injecting a tracer

into a stream and noting the time required for the radioactive pulse to travel the distance between two detectors or between the injection point and a single detector. Rapid injection and accurate timing may be avoided where the amount of radioactivity injected and the detection efficiency of a downstream detector are accurately calibrated. The total signal from the tracer is inversely proportional to the velocity with which it goes past the detector. The techniques have been used in pilot plants, refineries, chemical plants, and even to study the flow rates and patterns of rivers and ocean currents.

The disposition of materials in various process units has been evaluated. Stream splitting, recycling, residence times, entrainment in distillations, mixing, and unit inventories have been measured.

The techniques of use in industrial fluid processes also apply in other fields. Stream splitting in capillary gas chromatography sampling units has been studied. Blood flow rate, total blood volume, and heart function are examples of medical applications.

H. R. LUKENS, JR.

References

- Overman, R. T., and Clark, H. M., "Radioisotope Techniques," New York, McGraw-Hill Book Co., 1960. Describes general laboratory radioisotope procedures.*
- Kohl, J., Zentner, R. D., and Lukens, H. R., "Radioisotopes Applications Engineering," Princeton, N.J. D. Van Nostrand, 1961. Tracer applications are treated in detail, including tracer selection, calculations, and measurement.*
- Seymour Rothchild, Ed., "Advances in Tracer Methodology," Vol. 1, New York, Plenum Press, 1963. Places particular emphasis on the uses of tritium (H^3) and carbon 14.

Cross-references: ACCELERATORS, PARTICLE; ISOTOPES; NUCLEAR INSTRUMENTS; NUCLEAR RADIATION; NUCLEAR REACTORS; RADIATION CHEMISTRY; RADIATION, IONIZING, BASIC INTERACTIONS; RADIOACTIVITY.

RADIOACTIVITY

Radioactivity is the term applied to the spontaneous disintegration of atomic nuclei. It was one of the first and most important phenomena which led to our present understanding of nuclear structure. Credit for the discovery of radioactivity is usually given to Henri Becquerel, who made the observation in 1896 that penetrating radiation was given off by certain compounds of heavy elements in the absence of any external stimulus. Many other scientists were working in the field of radiation, however, and the announcement by Becquerel led to a flood of discoveries about the nature of the radiations which were emitted. A

* Both books contain abundant references to other works, general and specialized, concerning radioactive tracers.

number of workers determined that certain of the radiations from these radioactive substances could be deflected in a magnetic field, and by 1900 three separate types of rays were identified. They were given the names alpha, beta, and gamma rays.

Distinction is frequently made between natural radioactivity, which was observed by the early workers, and artificial radioactivity which was first produced by Joliot-Curie and Joliot in 1934. These workers bombarded Al^{27} with alpha particles to produce P^{30} . For purposes of our discussion, we shall not distinguish between the sources of the radioactive material.

A discussion of radioactivity requires that mention be made of the stability of nuclei. Stable nuclear species or nuclides exist for all elements having proton numbers in the range from 1 to 83 except for elements 43 and 61 (technetium and promethium). In general, elements having even atomic numbers have two or more stable isotopes, whereas odd-numbered nuclei never have more than two.

The assumption is usually made that all possible nuclides were formed in the original atomic production processes, and that those which remain at the present time do so because of some inherent stability. In general, this stability involves the neutron-proton ratio, and a number of theoretical studies have been undertaken to determine the conditions for the maximum stability for nuclei.

Stability may be considered from three different standpoints: relative to the size and the number of particles in the nucleus, the ratio of the neutrons and protons in the nucleus, and the ratio of the total mass-energy of the nucleus. A nucleus which is unstable with respect to its size will emit alpha particles whereas a nucleus unstable with respect to its neutron-proton ratio may emit a negative or positive electron or may capture an electron. If a nucleus is unstable with respect to its total energy, the excess energy may be given off as gamma radiation which is electromagnetic in nature. Let us consider these emissions in more detail.

Alpha Emission. It is evident that there are two types of forces existing in the nucleus. The first is a disruptive force arising from the repulsion of similarly charged particles. In addition, however, there are very strong attractive forces arising from the interactions of the nucleons. These attractive forces are very strong within the nucleus, but drop off quite sharply beyond about 10^{-12} cm from the center of the nucleus. Alpha particles, consisting of two protons and two neutrons, are, with certain exceptions, observed to come only from the larger nuclei. The mechanism of the emission process, however, is not a simple force phenomenon.

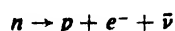
If we make calculations involving only the energies of the nucleus and the ejected particle, we would find, for example, that an alpha particle should come from a Ra^{226} nucleus with approximately 27 Mev of energy. Instead, the emerging particle is observed to have about 5.3 Mev. This difference in energy cannot be explained using

classical energy computations. The explanation of the alpha emission is usually given in terms of a "tunneling" effect, which is a quantum mechanical description first developed by Gamow and by Gurney and Condon. There are two equivalent ways of looking at this effect. One is to consider the alpha particle as being in motion inside the nucleus. In this picture, we visualize the particles as striking the potential "wall" of the nucleus. According to the quantum mechanical treatment for a particle striking such a barrier, there is a finite probability of passing through the barrier and appearing on the outside. Calculations making use of this probability give correct values for the observed half life and energies of the alpha particles from radioactive nuclides.

The other description of this process considers the alpha particle as being a wave packet with a very high probability of being found within the nuclear radius, but also having a finite probability of being outside the nucleus. According to this notion, the probability of finding the alpha wave packet at a distance greater than the nuclear radius, likewise gives the proper lifetime and energy values. It can be shown that this type of radioactivity is more probable in elements of high atomic number, and present extrapolations for this type of instability suggests that the highest atomic number element which can exist may be in the region of 108 or 109. It appears that elements 105 and higher will not exist long enough to be identified chemically, since the alpha emission will be extremely probable, affording an extremely short half-life.

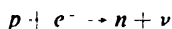
Beta Radiation. There are over 400 nuclear species which have been artificially produced in the laboratory. Nearly all of these have been produced by the reactions which give rise to a net gain or loss of neutrons from the stable nuclei. In such a case the residual nucleus is characterized by having a neutron-proton ratio higher or lower than stable nuclei of that element. Similarly, if one considers alpha-emitting nuclei as being the "stable" nuclides for elements above lead, certain nuclides may be formed after the alpha emission with neutron-proton ratios higher or lower than the original element. The general decay process for nuclides with differing n/p ratios involves the reorganization of the nucleus in a manner which will leave a nucleus having a neutron-proton ratio corresponding to that requisite for stability.

Let us first consider the case in which a nucleus has a neutron-proton ratio higher than a stable isotope of that element. For example, the nuclei of all stable phosphorous atoms contain 15 protons and 16 neutrons. If a nuclear reaction takes place which leaves a nucleus with 15 protons and 17 neutrons we have a nucleus of P^{32} . Since only P^{31} occurs in nature, we know that some adjustment will take place to bring about stability. In this case, one of the neutrons is transformed into a proton, a negative electron, and a neutrino (specifically an anti-neutrino) which is shown in the following reaction:



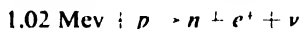
The negative electron cannot exist as part of the nucleus and is ejected from the nucleus along with the neutrino. When the electron is investigated, it is found to be identical to other negative electrons, but when it is formed in this process, it is given the name negatron or negative beta particle. The residual nucleus then contains 16 protons and 16 neutrons. The element possessing 16 protons is sulfur, so the nucleus resulting from this negative beta emission is S^{32} . In general, then, nuclei with neutron-proton ratios higher than stable nuclei eject a negative electron from the nucleus and are thus transmuted into a nucleus of the next higher atomic number.

There are also types of nuclear reactions which may leave a nucleus with a neutron-proton ratio lower than the corresponding stable nucleus. In this type of instability, there are two processes which may compete with one another for the production of a stable nucleus. In the first of these, an extranuclear electron may be captured by the nucleus and combined with one of the protons. The most likely electron taking part in this process is one from the K level. L or M level electrons may be "captured," however. This electron capture reaction may be shown as follows:



A neutrino is also ejected in this process. When this reaction takes place, the nucleus then contains one less proton than before, and is thus a nucleus of one lower atomic number. For example, if Fe^{55} "decays" by electron capture, the resulting nucleus of Mn^{55} is stable. This process can take place whenever the neutron-proton ratio is too low for stability.

The reaction which competes with electron capture may occur if the neutron-proton ratio is low, and if there exists a certain minimum mass-energy difference between the unstable nucleus and a possible stable nucleus having one less proton. If there is at least 1.02 Mev mass-energy difference a proton may transform into a neutron, a positive electron (or positron), and a neutrino according to the following equation:



The positron is then ejected from the nucleus. The characteristics of the positron are identical to those of the negatron except for its positive charge. A number of cases are known in which both electron capture and positron emission processes take place in the same nuclear species. A radioactive nuclide is characterized as having a certain "branching ratio" when more than one type of decay is possible.

Gamma Radiation. The usual modes of decay which involve the re-organization of the nucleus are those described above. In many cases, however, another step is involved in attaining final stability. After one of the nuclear transformations described above takes place, the nucleus may still possess excess energy. In this case, the extra energy is given off directly as gamma radiation. These are electromagnetic radiations which have energies corresponding to the difference in energy

levels in the nucleus from which they come. A particular nuclide thus exhibits a certain pattern or disintegration scheme by which it decays. For example, P^{32} decays by negative beta emissions, which are not followed by gammas. On the other hand, Co^{60} emits negative beta radiation which is followed in each case by two gamma rays in cascade. These gamma radiations have energy of 1.17 and 1.33 Mev respectively. In general, the gamma radiations are emitted in time periods less than 10^{-12} seconds following the first transmutation step. In some cases, however, excited energy states may exist for significantly longer periods. Experimental determinations of the lifetimes of these slower gamma ray transitions range from 10^{-8} second to several months. If such an excited state exists in a nuclide for a period long enough to be measured experimentally, the nuclide is called a nuclear isomer, and the transition process involving such gamma radiation is called an isomeric transition (I.T.).

It should be mentioned also that gamma radiation is given off following nuclear reactions. Such gamma rays called "capture" or "prompt" are discussed in conjunction with nuclear reactions, although isomeric states are often formed in this manner, and isomeric transitions may leave the nuclei in radioactive rather than stable states.

Decay Schemes and Units. Information as to the radioactive transitions which take place in a given case are frequently presented in what is called a decay scheme. A group of simple decay schemes is shown in Figure 1. These indicate the type of radioactive process which the nucleus undergoes, the energies of the radiations given off, and the branching which may take place. Most nuclei have decay schemes which are much more complex than those represented in the Figure, but they involve only multiple occurrence of the processes which have been described.

Decay Rate. It can be seen from the foregoing that a nucleus may change from an unstable to a stable form by one of several decay processes. We can thus speak of the decay rate of a sample of radioactive material in disintegrations per unit time—usually disintegrations per second. This refers specifically to the transformation of the nucleus, and does not give any indication of the kind or energy of the radiations emitted. The unit of activity is the "curie" and is defined to be 3.7000×10^{10} disintegrations per second. Submultiples or multiples of this unit in common use are micro-, milli-, kilo-, and megacurie. A sample containing one millicurie of radioactive material is a sample which decays at the rate of 3.7000×10^7 disintegrations per second.

We have no way to determine when any given nucleus will decay, but some of the most important work in the early study of radioactivity involved the study of decay rates. It was shown very early that the rate of radioactive decay is proportional to the amount of the radioactive material present. This can be expressed in the following equation:

$$A = -\lambda N$$

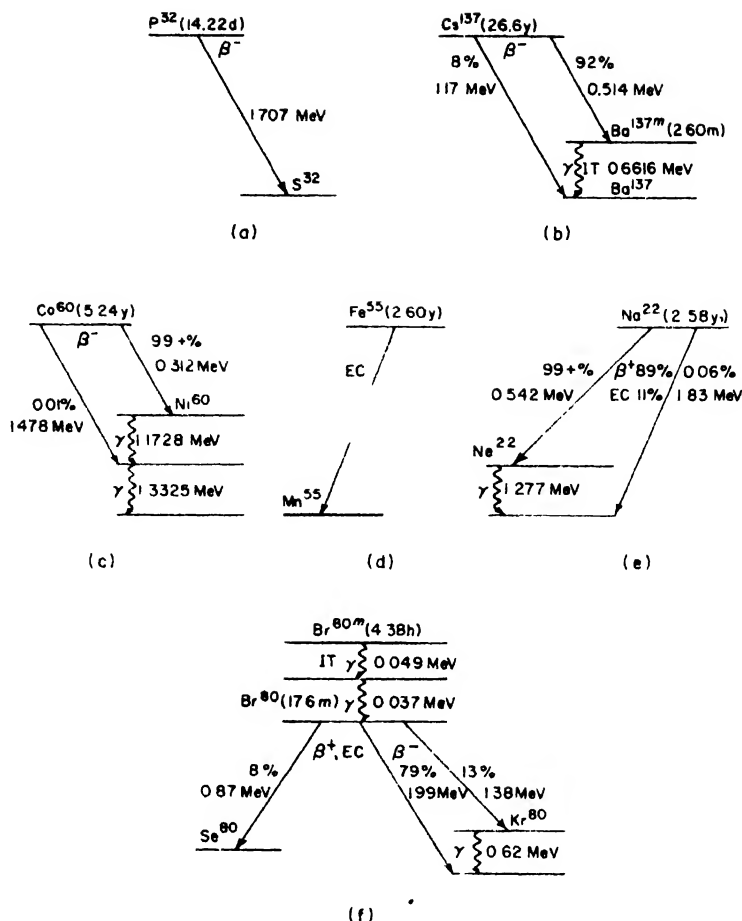


FIG. 1. Decay schemes.

in which A is the disintegration rate, N is the number of radioactive atoms present, and λ is a proportionality constant characteristic of each radioactive species. This relationship is sometimes important, but often a more useful relationship is the following:

$$N = N_0 e^{-\lambda t}$$

in which N_0 is the number of atoms present at some reference time, and N is the number of atoms present at some time t later, and e is the base of the natural logarithms. This exponential decay relationship is of fundamental concern in working with radioactive materials, even though it is valid only for a statistically large number of atoms (i.e., all unstable nuclei will eventually decay).

The above equations are valid for any single radioactive species, and the exponential expression states that a given fraction of nuclei will decay in a given time period. In practice, we frequently refer to the time required for one-half of the atoms to decay. This is called the half-life of the radioactive material. It is related to the

decay constant in the following way:

$$T_{1/2} = \frac{0.693}{\lambda}$$

It can also be shown that the average life, $\tau = 1.44$ times the half-life.

There are many cases in which a series of radioactive steps take place. For example, U^{238} undergoes 14 successive decays before arriving at the stable end product, Pb^{206} . These steps involve the emission of successive alpha and beta rays along with gamma radiation in some cases. Although the calculations are somewhat more complex, it is possible to determine the disintegration rate of each of the daughter radioactive species formed in these processes. Details of this type of calculations are given in the references below.

RALPH T. OVERMAN

References

Evans, R., "The Atomic Nucleus," New York, McGraw-Hill Book Co., 1955.

Glasstone, S., "Sourcebook on Atomic Energy," Second edition, Princeton, N.J., D. Van Nostrand Co., 1958.

Lapp and Andrews, "Nuclear Radiation Physics," Third edition, Englewood Cliffs, N.J., Prentice-Hall, 1963.

Overman, R. T. and Clark, "Radioisotope Techniques," New York, McGraw Hill Book Co., 1960.

Cross-references: ELECTRON, NEUTRINO, NEUTRON, NUCLEAR RADIATION, NUCLEAR STRUCTURE, NUCLEONICS, POSITRON, PROTON, TUNNELING.

RAMAN EFFECT AND RAMAN SPECTROSCOPY

The Raman effect is the phenomenon of light scattering from a material medium, whereby the light undergoes a wavelength change in the scattering process. For a given medium, the Raman scattering per unit volume is of the order of one-thousandth of the intensity of the ordinary or Rayleigh scattering, in which there is no change of wavelength (see LIGHT SCATTERING). The Raman-scattered light bears no phase relationship with the incident light, whereas the Rayleigh light is a residual effect resulting from the departure of the incident and the scattered light from complete mutual coherence. The Raman intensity per molecule is thus independent of the state of the medium, apart from a certain small refractive index effect. The Rayleigh intensity per molecule, on the other hand, depends strongly on the degree of randomness of the spatial positions and orientations of the molecules of the medium; it is small for a crystal at the absolute zero and greatest for a gas at low density.

Scattering of light with change of wavelength was predicted in 1923 by Smekal, inspired by the discovery of the Compton effect. The Raman effect was discovered experimentally in 1928 in India by Raman and Krishnan, who showed that the spectrum of the scattered light of liquids and solids, strongly illuminated with monochromatic light, contains frequencies which are not present in the exciting light and which are characteristic of the scattering medium. Independently, and almost simultaneously, Landsberg and Mandelstam in Russia discovered the effect in crystals. The phenomenon is called combination scattering of light in present-day Russian literature.

The method of observing the Raman spectrum is usually some modification of the arrangement introduced originally by R. W. Wood. The specimen, e.g., a liquid contained in a tube 1 cm in diameter and 10 cm long, is strongly illuminated along its length by mercury arcs, with filters interposed between the arcs and the tube to isolate monochromatic radiation if necessary. The scattered light is observed along the axis of the tube through a plane window at one end of the tube. The other end of the tube is usually drawn out in a cone which, when blackened, forms a dark background against which the

scattered light is observed. The scattered light is analyzed by a spectrograph or recording spectrometer.¹

The experimentally confirmed laws of Raman scattering are as follows:

(a) The pattern of Raman lines, expressed as frequency shifts from the exciting line, $\Delta\nu_i$ ($i = 1, 2, \dots$), is independent of the exciting frequency.

(b) The pattern of Raman frequency shifts, $\Delta\nu_i$, is symmetrical about the exciting line. However, the lines on the low-frequency side of the exciting line (Stokes lines) are always more intense than the corresponding lines on the high-frequency side (anti-Stokes lines). The ratio of the intensities of corresponding anti-Stokes and Stokes lines is $I_a/I_s = \exp(-\Delta\nu_i hc/kT)$, where the Raman shift $\Delta\nu_i$ is expressed as usual in cm^{-1} . Thus anti-Stokes lines for $\Delta\nu_i > \sim 1000 \text{ cm}^{-1}$ are too weak to be observed at room temperature.

(c) A given Raman line shows a degree of polarization which depends on the origin of the line and on the experimental arrangement. For strictly transverse observation, i.e., observation at right angles to the incident light, the depolarization factor, ρ_n , has a value in the range 0 to 6/7 for unpolarized incident light.

Fig. 1 shows a schematic diagram of the Raman spectrum of carbon tetrachloride. The four Raman shifts, $\Delta\nu_i$ ($i = 1, 2, 3, 4$), are 218, 314, 459 and 775 cm^{-1} ; the corresponding values of ρ_n are 6/7, 6/7, 0, and 6/7.

The Raman shifts, $\Delta\nu_i$, correspond to energy differences (in cm^{-1}) between discrete stationary states of the scattering system. Thus, in the quantum picture, the incident photons collide *elastically* with the molecules to give Rayleigh scattering, or *inelastically* to give Raman scattering, the latter process being much less probable than the former. For a Stokes Raman line, the photon furnishes energy to raise the molecule from a lower to a higher state; for an anti-Stokes line, the molecule must furnish energy to the scattered photon and move to a lower energy state. The anti-Stokes line thus originates in a less highly populated state and is weaker than the corresponding Stokes line. The Raman process can be described classically, but not as accurately, as the modulation of the scattered light wave by the internal motions of the scattering molecule, the Raman lines constituting "sidebands" of the Rayleigh "carrier" frequency.

A rigorous theory of the Raman effect, based on the quantum theory of dispersion, has been given by Placzek.² The process of light scattering can be visualized as the absorption of an incident photon of frequency, ν_0 , by a molecule in a given initial state, thus raising the molecule to a "virtual" state from which it immediately returns to a final state emitting the scattering photon.

Fig. 2 illustrates the production of the Rayleigh line and Stokes and anti-Stokes Raman lines for a two-level system. If the frequency of the exciting light is such that the virtual state coincides with a real energy state of the system, the Raman effect goes over into molecular fluorescence provided

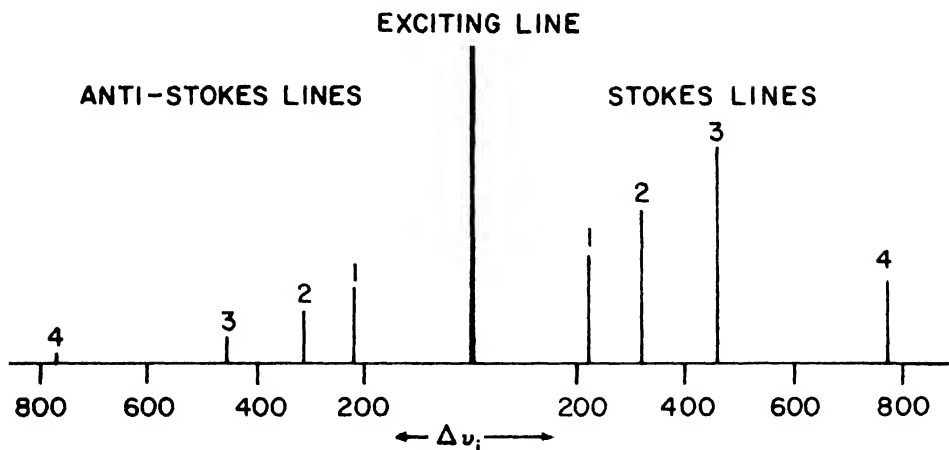


FIG. 1. Schematic diagram of the Raman spectrum of carbon tetrachloride. The heights of the Raman lines represent their relative intensities.

the states involved can combine with one another according to the selection rules.

In principle, the Raman shifts, $\Delta\nu_i$, can represent energy differences between electronic, vibrational or rotational energy states. In practice, Raman spectroscopy has been concerned chiefly with the determination of vibrational frequencies of polyatomic molecules from the scattered light spectrum and the correlation of the observed frequencies with possible modes of vibration of the molecules. In this role, Raman spectroscopy forms an important complement to near-infrared spectroscopy which also furnishes precise information on molecular vibration frequencies.³

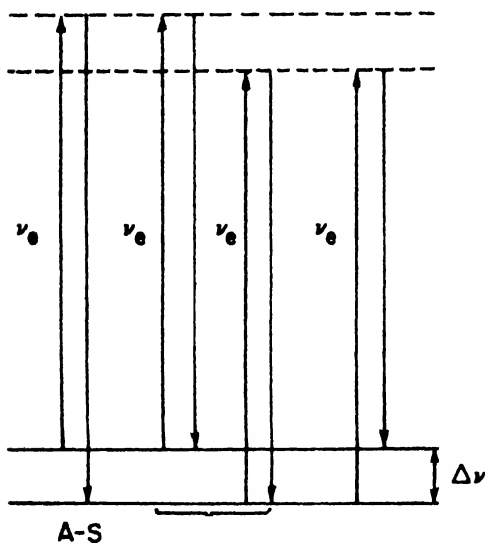


FIG. 2. Illustrating the production of Stokes (S) and anti-Stokes (A-S) Raman lines and the Rayleigh line (R) by an exciting frequency, ν_e , for a two-level system. The dashed lines represent virtual states of the system.

The so-called polarizability theory of Placzek² is an approximate theory which is particularly useful in relating observed vibrational Raman lines to the normal vibrations of the scattering molecule. In the equation, $m = \alpha E$, m is the electric dipole induced in the molecule by the oscillating light field E . The constant of proportionality, α , is the molecular polarizability, which can be developed in a Taylor series in the vibrational coordinate q ,

$$\alpha = \alpha_0 + (\partial\alpha/\partial q)_0 q + (1/2)(\partial^2\alpha/\partial q^2)_0 q^2 + \dots$$

The intensity of the fundamental Raman band, corresponding to the $v = 0$ to $v = 1$ transition where v is the quantum number of the vibration, depends essentially on the rate of change of polarizability, $\alpha' \equiv (\partial\alpha/\partial q)_0$, with respect to the coordinate q . The vibration is active in Raman scattering if, because of the symmetry properties of the vibration, α' is not identically zero. Also the depolarization, ρ_n , of a Raman band can be related by the polarizability theory to the symmetry of the vibration. Thus, the single polarized ($\rho_n = 0$) line of carbon tetrachloride at 459 cm^{-1} (Fig. 1) can be identified immediately with the "breathing" vibration of the CCl_4 molecule, and the three depolarized ($\rho_n = 6/7$) lines with the three degenerate vibration forms of the molecule.

Since the infrared activity of a molecule depends on the molecular dipole, whereas the Raman activity depends on the molecular polarizability, the selection rules for two types of spectra can be very different. An extreme case occurs for molecules with a center of symmetry: vibrations which are active in infrared absorption are not active in Raman effect and vice versa.

Although most investigations of the Raman effect have dealt with liquids and solids, i.e., high-density materials, systematic Raman studies of low-pressure gases have recently been made possible by the development of special equipment.^{4,5} Since the molecules in gases are freely

rotating, pure rotational Raman spectra and the rotational structure of vibrational bands can be observed. Here again, Raman studies supplement infrared investigations since the rotational selection rules are $\Delta J = 0, \pm 1, \pm 2$ for the former and $\Delta J = 0, \pm 1$ for the latter, where ΔJ is the allowed change in the rotational quantum number J .

The recent development of optical masers or lasers has great importance for the future of Raman research. It is probable that lasers will replace mercury arcs for excitation of Raman spectra. In addition, two new phenomena have been discovered which will undoubtedly enlarge the scope of Raman investigations. The first of these is the "stimulated" Raman effect,⁶ a coherent and directionally dependent form of Raman scattering, which results from the intense coherent electric fields set up in a medium through which laser radiation is passing. The second is the "inverse" Raman effect,⁷ in which, under certain conditions, the Raman lines appear as absorption lines on a continuous background.

H. L. WELSH

References

1. Harrison, G. R., Lord, R. C., and Loofbourow, J. R., "Practical Spectroscopy," Edgewood Cliffs, N. J., Prentice-Hall, 1948.
2. Placzek, G., *Handbuch der Radiologie*, 6, No. 2 (1934).
3. Herzberg, G., "Infrared and Raman Spectra of Polyatomic Molecules," New York, Van Nostrand 1945.
4. Welsh, H. L., Stansbury, E. J., Romanko, J., and Feldman, T., *J. Opt. Soc. Am.*, 45, 338 (1955).
5. Stoicheff, B. P., *Advan. Spectry.*, 1, 91 (1959).
6. Hellworth, R. W., *Phys. Rev.*, 130, 1850 (1963).
7. Jones, W. J., and Stoicheff, B. P., *Phys. Rev. Letters*, 13, 657 (1964).

Cross-references: LASER, LIGHT SCATTERING, POLARIZED LIGHT.

RARE EARTHS

The rare-earth elements (also called "rare earths", "lanthanides" and "lanthanons") are a group of fifteen elements of similar properties in Group III of the periodic table. Their properties are somewhat similar to those of scandium, yttrium, and actinium in the same group. The rare earths have atomic numbers 57 to 71, inclusive, and are, in serial order: lanthanum (La), cerium (Ce), praseodymium (Pr), neodymium (Nd), promethium (Pm), samarium (Sm), europium (Eu), gadolinium (Gd), terbium (Tb), dysprosium (Dy), holmium (Ho), erbium (Er), thulium (Tm), ytterbium (Yb), and lutetium (Lu). The rare earths from lanthanum to samarium are called the "cerium earths"; those from europium to dysprosium are "terbium earths." Those from holmium to lutetium are called "yttrium earths" because of their resemblance to yttrium which is similar to and always occurs with the rare earths.

The rare earths are inner transition elements characterized by progressive filling up of the $4f$ electrons without changing the outer $5s^2 5p^6 5d^0$ or $6s^2$ levels, resulting in a concurrent decrease in atomic size. This reduction in atomic radii with increasing atomic number is known as the *lanthanide contraction*. In this respect, and also in chemical behavior, the rare earths resemble the actinides, the second rare-earth-like series beginning with actinium, which shows a similar *actinide contraction* (see TRANSURANIUM ELEMENTS).

Occurrence. The only commercial minerals of importance are monazite and monazite sand, and bastnasite. Monazite is a rare-earth-thorium phosphate which occurs as small crystals in many acid granites and is recovered from placers resulting from the weathering of such rocks. It also occurs in some pegmatites, but commercial pegmatite deposits are rare.

Bastnasite is a rare-earth fluorocarbonate which occurs in hydrothermal deposits associated with barite, fluorite, and calcite.

Important monazite deposits are placers in Idaho, beach sands in Florida, Brazil, and India, and massive pegmatite monazite in the Union of South Africa. Bastnasite deposits are found in California and New Mexico, but only the California deposit is important.

Other minerals, sometimes useful for the extraction of the heavier rare earths, are gadolinite, $\text{Fe}(\text{RE})_2\text{Be}_2(\text{Si}_2\text{O}_{10})$; samarskite, a rare-earth-iron-calcium-magnesium niobotantalate; and xenotime, an yttrium-earth-rare-earth phosphate.

The cerium earths predominate in monazite and bastnasite. Cerium is the most abundant rare earth, being comparable to boron in occurrence in the earth's crust. Promethium is found only in the products from uranium fission.

Properties. The elements are trivalent in most compounds. Cerium, praseodymium, and terbium exist in the tetravalent state, and samarium, europium, and ytterbium form easily oxidized divalent compounds. Due to similarities in atomic structure, many chemical properties of the rare earths are quite similar and vary only slightly from one rare earth to the adjacent neighbor. Consequently, their separation is difficult unless use is made of oxidation states other than the trivalent. With increasing atomic number, the rare earths become less basic, the salts generally become more soluble, and the differences between adjacent members decrease.

Ordinary rare-earth similarities are usually based on conventional chemical observations. Thus, in aqueous systems, the relatively large degree of hydration of rare-earth ions tends to mask differences. In non-solvent systems where the masking effect is not present, differences between rare earths are more apparent. Rare-earth metals and simple compounds do indeed show more marked differences, and the more sophisticated research on rare earths since about 1950 shows that mutual similarity is not necessarily an inherent property of these elements.

With the exception of lanthanum and lutetium, rare earth compounds show characteristic sharp

absorption bands in the ultraviolet and visible spectra. This absorption is responsible for the pastel colors of the colored rare earth salts (green Pr, pink Nd, yellow Sm and Ho, rose Er, and pale green Tm).

Transitions in the 4f electrons account for most unusual properties: sharp absorption spectra in the ultraviolet, visible, and infrared—useful in optical devices, in the study of spectra theories, and in lasers. Complete pairing or formation of complete inner electron shells, resulting in balancing of magnetic moments causes ions to be diamagnetic (as for La, Lu, and Y); unpaired electrons lead to paramagnetic ions (as for Dy, Tb, and Ho).

Rare-earth Metals. The metals have a silver-gray luster which tarnishes quickly in air if the metals are easily oxidized (La, Ce, Sm, Eu). Hardness varies from that comparable to lead (Eu, Yb), to tin (Ce, La), to zinc (Nd, Pr), to mild steel (Gd, Y). Densities (grams per cubic centimeter) vary from 6.18 for La and 6.8 for Ce to 9.8 for Tm and 7.0 for Yb. Melting points vary from 804°C for Ce to about 1700°C for Lu, and the boiling points are in the range 1490 to 4200°C.

Being active reducing agents, the metals react slowly with water and are soluble in dilute acids. They are pyrophoric when finely divided, cerium igniting in air at 150 to 180°C. Above 200°C, they combine directly with halogens, and they form nitrides with nitrogen above 1000°C. Interstitial hydrides approximating RH_2 or RH_3 are formed by absorption of hydrogen.

The mixture of rare-earth metals made commercially without appreciable separation of rare earths is known as "misch" metal. It contains about 22 per cent La, 50 per cent Ce, 18 per cent Nd, 5 per cent Pr, 1 per cent Sm, and 2 per cent other rare-earth metals. It is often sold as "cerium" metal.

Rare-earth Compounds. Common water-soluble salts are the acetates, chlorides, nitrates, and sulfates. The carbonates, oxalates, hydroxides, oxides, phosphates, and fluorides are insoluble. Cerous (Ce^{+3}) salts are similar to the other trivalent rare earth salts, while ceric (Ce^{+4}) salts are more like those of thorium. Trivalent acetates and sulfates show decreased solubility in hot solutions.

Acetates are made by treating the hydroxide, carbonate or oxide with acetic acid. Carbonates are precipitated by the addition of alkali carbonates to neutral rare-earth solutions; they are only slightly soluble in excess of alkali carbonate. Fluorides are precipitated in hydrated form on adding hydrofluoric acid or soluble fluorides to rare-earth solutions. The precipitates are insoluble in excess hydrofluoric acid and in mineral acids. Anhydrous fluorides are made by hydrofluorinating the oxides at elevated temperatures.

Chlorides, bromides and iodides are prepared by dissolution of the hydroxide, oxide, or carbonate in the halogen acid. The salts crystallize from water as hydrates. Anhydrous chlorides and bromides are made by heating oxides with the

ammonium halide, followed by sublimation of the excess ammonium salts in vacuum. Nitrates are formed similarly to the other water-soluble salts. They form many double nitrates with alkali and alkaline earth nitrates. Oxalates are precipitated from slightly acid solution with oxalic acid or alkali oxalates. They are important in the analysis for rare earths, and in the separation of rare earths from other metals. Hydroxides are precipitated from solution by alkali and ammonium hydroxides. The sulfates are characterized by their formation of sparingly soluble double sulfates with alkali sulfates. The most important double sulfate is the sodium salt, $RE_2(SO_4)_3 \cdot Na_2SO_4 \cdot 2H_2O$.

Extraction and Separation. The extraction of rare earths from monazite is commercially important. The ore is opened by heating with sulfuric acid to form anhydrous rare-earth and thorium sulfates and phosphoric acid, or by heating with sodium hydroxide solutions to form rare-earth and thorium hydroxides and sodium phosphate. The reaction products are lixiviated in water, and if the alkaline method is used, the washed rare-earth-thorium hydroxides are dissolved with acid.

Thorium is separated from the rare earths by fractional basicity precipitation or by precipitation of compounds such as thorium pyrophosphate. Rare earths are usually recovered from the thorium filtrates by precipitation of the double rare-earth sodium sulfate. The double sulfate precipitate is converted to rare-earth hydroxide which serves as the starting material for making commercial rare-earth salts.

Cerium is separated from the rare earths by oxidation to the tetravalent ceric state, followed by basicity separations or crystallizations of insoluble ceric compounds. Ceric salts are generally much less soluble than those of the trivalent rare earths. Crystallization of ammonium hexanitratocerate or ammonium trisulfatocerate, precipitation of basic ceric nitrates or sulfates, and fractional basicity separations with the hydrous oxides are commonly used procedures to separate and purify cerium.

Separation of the cerium-free rare earth mixture (often called "didymium") formerly was done by long series of fractional crystallizations and fractional precipitations. Ion exchange and solvent extraction methods of separation have replaced tedious fractional crystallization, and are used commercially to produce both technical and ultrahigh-purity rare earths. Exclusive of non-rare earth impurities, the purity of commercially available individual rare-earth compounds is usually 99.9 per cent, and in some cases as high as 99.9999 per cent.

Uses. Most uses are based on unseparated rare-earth mixtures, or on technical cerium and "didymium" materials. About one-fourth of the rare-earth salts produced are used in carbon arc lighting. Rare-earth-cored carbons are indispensable to the motion picture industry and are also used in military searchlights.

Rare-earth metal (misch metal) and cerium

metal are important in the manufacture of lighter flints and certain alloys. Rare-earth-zirconium-magnesium alloys have outstanding high-temperature properties. Certain types of ferrous and stainless alloys are improved by the addition of misch metal or rare-earth compounds.

Rare-earth salts have important uses in the coloring and decolorizing of glass. Specially prepared cerium oxide and rare-earth oxide are widely used in polishing spectacle and optical instrument lenses, as well as mirrors, glass products, and granite.

Miscellaneous uses of the rare earths are: lanthanum oxide in silica-free optical glass, neodymium oxide as a coloring material for novelty glassware, "didymium" salts in temperature-compensating ceramic capacitors, rare-earth oxalate as a nausea preventive, rare-earth compounds as activators in phosphors, and various rare-earth salts in catalyst manufacture. Cerium compounds are used as laboratory reagents and as scavengers in explosive manufacturing.

As nuclear poisons, Sm, Gd, Eu, and Dy materials are useful because of the very high neutron cross sections of some of their isotopes.

Yttrium oxide is a component of microwave ferrites of the garnet type. Most of the individual rare earths are important as host materials or dopants in laser crystals.

HOWARD E. KREMFERS

References

- Gschneidner, K. A., "Rare Earth Alloys," Princeton, N. J., D. Van Nostrand Co., Inc., 1961.
 Tipton, C. R., Ed., "Reactor Handbook," Second edition, Vol. I, New York, Interscience Publishers, Inc., 1960.
 Spedding, F. H., and Daane, A. H., "The Rare Earths," New York, John Wiley & Sons, Inc., 1961.
 Pascal, Paul, "Nouveau Traité de Chimie Minérale," Vol. VII, Paris, Masson et Cie, 1959.

Cross-references: ELEMENTS, TRANSURANIUM ELEMENTS.

REACTOR. *See* NUCLEAR REACTORS.

REACTOR SHIELDING

Currently, nuclear reactors utilize the FISSION process to "burn" uranium or plutonium fuels. In this process, radiations are emitted which must be intercepted to prevent them from injuring nearby personnel or equipment. The material which does so is called shielding. In the FUSION process, which may some day be used in another type of nuclear reactor, energy is released by charged-particle reactions, principally those involving deuterium or tritium. The radiations from these reactions are not as intense as those from the fission process. In order to solve a shielding problem, it is required to know the sources of radiation, the mechanisms for interaction of this radiation with matter, the tolerable levels of radiation, and the most efficient and precise methods of calculating the attenuation of that

radiation on its passage through a radiation shield.

Sources of Radiation. Prompt Neutrons. In a fission event, the fission fragments are produced in a highly excited state so that they boil off neutrons as they travel away from the site of the fission. These neutrons, called prompt neutrons, are distributed in energy according to a formula due to Cranberg *et al.*:¹

$$N(E) = 0.453e^{-E/0.965} \sinh(2.29E)^{1/2}$$

where E is the neutron energy in million electron volts and $N(E)$ is the fraction of all prompt neutrons produced per unit energy interval about E . The most probable prompt neutron energy is $\frac{1}{2}$ MeV and the mean energy is $2\frac{1}{2}$ MeV. In the range of 4 to 14 MeV, the fission spectrum can be approximated by a simple exponential form also due to Cranberg *et al.*:

$$N(E) = 1.75e^{-0.766E}$$

Note that for high energies, the yield decreases strongly with increasing energy.

Delayed Neutrons. In addition to the neutrons released instantaneously as described in the previous section, neutrons are released in the radioactive decay of some of the fission products. These neutrons are given off from a set of isotopes which decay with half-lives varying from 0.18 to 54.5 seconds. The fraction of all neutrons produced by the fission process which are "delayed" is 0.0029 for U-233, 0.0067 for U-235, and 0.0021 for Pu-239.² Since in addition they are of lower energy than fission neutrons, they are not usually of great consequence in reactor shielding. In the case of circulating-fuel reactors such as the Molten Salt Reactor Experiment at Oak Ridge, however, they are of great importance since their presence requires that the heat exchanger be shielded as a neutron source. Furthermore, activation of the coolant by these neutrons must be considered.

Prompt Fission Gamma Rays. At the instant of fission, one or more gamma rays are given off immediately, having on the average a total energy of about 7.2 MeV per fission event, the photons being distributed in energy approximately according to the following exponential:³

$$\Gamma_p(E) = 8.0e^{-1.10E} \text{ MeV}^{-1}$$

This formula fits the data within 40 per cent between 1 and 7 MeV. The fit is only somewhat worse down to 0.4 MeV, below which the gamma rays have not been measured. Significant numbers of gamma rays are not observed above about 7.3 MeV.

Fission-product-decay Gamma Rays. Fission products are radioactive and most of them decay with gamma-ray emission. For times after fission less than about 200 seconds, the fission-product gamma-ray spectrum can be approximated by the formula:³

$$\Gamma_{fp}(E) = 6.0e^{-1.10E} \text{ MeV}^{-1}$$

For times greater than 200 seconds, the formula gives an overestimate of the region above 3.5 MeV.

Since most of the fission-product gamma rays are produced at short times after fission, however, the foregoing formula is adequate for an operating reactor, which means that for simplicity, the total of fission-product and prompt-fission gamma rays can be represented by a single formula:³

$$\Gamma(E) = 14e^{-1.10E} \text{ MeV}^{-1}$$

After a reactor has been shut down, fission-product gamma rays will continue to be given off. The total gamma-ray energy given off per second per fission at time t seconds after the fission event is given approximately by the Way-Wigner formula:⁴

$$\Gamma(E) = 1.5t^{-1.2} \text{ MeV/sec-fission}$$

where t is the time after fission in seconds.

Capture Gamma Rays. Neutrons captured in the reactor core, reflector, or shield produce capture gamma rays, which usually have a total energy of about 7 MeV per capture event. These usually comprise the single most important gamma-ray source, from the shielding viewpoint. Sometimes boron is added to a shield to suppress capture gamma rays by absorbing neutrons with emission of an alpha particle and (93 per cent of the time) one $\frac{1}{2}$ -MeV gamma photon.

Inelastic-scattering Gamma Rays. Neutrons of sufficiently high energy can be scattered inelastically so that the struck nucleus is left in an excited state from which it decays, usually at once, by gamma-ray emission. The gamma-ray photons so produced have total energy less than the kinetic energy of the neutron. They may be produced in heavy non-magic nuclei by both low- and high-energy neutrons, in light elements by high-energy neutrons only, and not at all in hydrogen.

Interaction Processes. Interaction processes are described in other sections in more generality. Here we discuss them from the point of view of shielding.

Neutron Interaction Processes. The three possible processes are elastic scattering, inelastic scattering, and absorption, the last including all processes in which neutrons become part of the nucleus and hence "disappear." Generally, elastic scattering is most frequent but least effective in attenuation; inelastic scattering is nearly as frequent for fast neutrons but much more effective; and absorption is very rare except for slow neutrons, but it is, of course, very effective.

Neutron total cross sections generally decrease with increasing neutron energy, so that the fastest neutrons are most penetrating. Since collisions slow neutrons down, succeeding collisions are more likely, so that penetration can be quite well described in terms of a first flight at the initial energy with a succeeding short migration. The "effective removal cross section," which is the equivalent pure absorption cross section that best describes the observed neutron attenuation, is equal to about $\frac{1}{3}$ of the total cross section

because of the reduced effectiveness of elastic scatterings in most elements. For hydrogen, $\frac{1}{3}$ is replaced by about $\frac{1}{2}$.

For a fission source, the more penetrating fast neutrons are also least populous at birth. This means that for thin shields the neutrons which penetrate are born at lower energies than for thick shields, which "filter out" the lower energies. For most reactor shields, the neutrons which penetrate are born at about 8 MeV. Figure 1 shows the effective removal as well as the total (at 8 MeV) cross sections for the gamut of elements. The macroscopic cross section, Σ , divided by density ρ , is shown; this quotient is independent of material density.

Gamma Interaction Processes. The three dominant processes are: (1) photoelectric effect, (2) Compton scattering, and (3) electron pair production. The cross sections for these three processes vary with atomic number, Z , approximately as Z^5 , Z and Z^2 . As for the variation of the cross sections with energy, those for the photoelectric effect are generally greatest for low energies; those for the Compton effect vary similarly, but not nearly so strongly; and those for pair production increase strongly with increasing energy. For shielding purposes, (1) and (3) may be treated as absorption whereas (2) is a scattering with energy reduction. It follows from the foregoing that elements of large Z , such as Pb, are best for low- (<1 MeV) and high- (>6 MeV) energy gamma rays. For low Z elements (such as those in ordinary concrete), the Compton effect dominates for all but the lowest-energy gamma rays and even for high energies it dominates at the minimum in the total cross section (~ 3.5 MeV in Pb). Since the Compton effect cross section varies proportionally to Z , which is roughly proportional to material weight, the effectiveness of a gamma-ray shield commonly varies proportionally to the product of its thickness and density. Hydrogen (for which Z/A is at least twice as large as for any other element) is a special case. It would be the best shield on a weight basis for medium gamma-ray energies if its low density were not important.

Attenuation Calculation. Calculating the attenuation of a shield consists in calculating the effect outside a shield of radiation emanating from a region inside. For a point gamma-ray source, dose rate is given simply by:

$$D(R, t, E) = \frac{SB(\mu t) e^{-\mu t}}{4\pi R^2} \cdot \int E$$

where

S = source strength, photons emitted per unit time,

R = separation distance between source and point at which flux is calculated,

μt = number of mean free paths (mfp) of the radiation on a straight-line path along R

$B(\mu t)$ = buildup factor, the ratio of dose rate due to all photons to that due to particles which have not been scattered in traversing the medium,

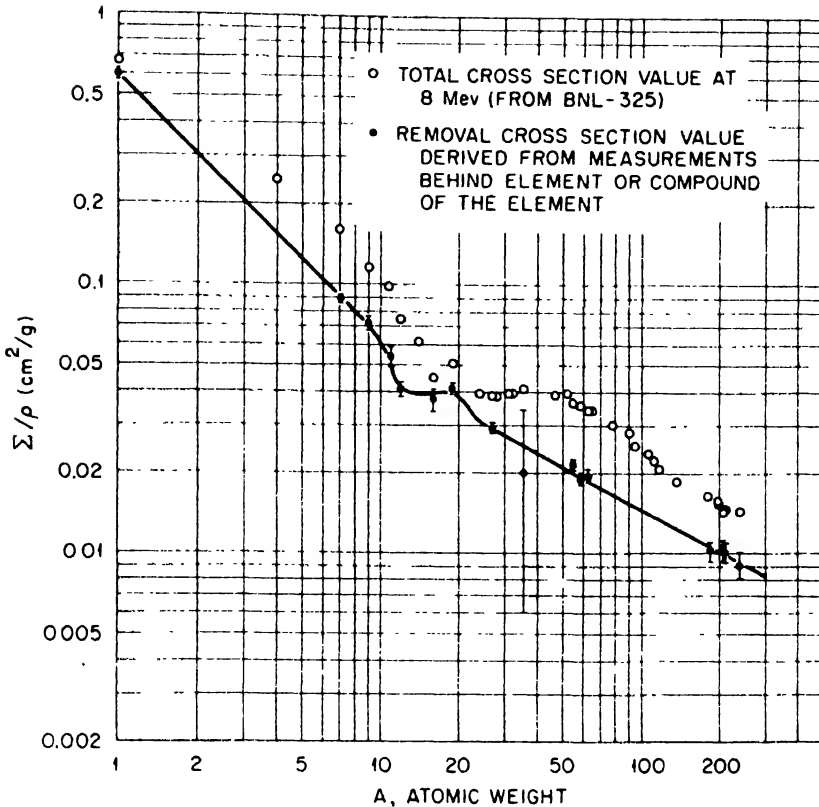


FIG. 1. Total and removal cross sections (in square centimeters per gram) for all elements vs their atomic weight.

f = the dose rate per unit energy flux of photons as emitted by the source,
 E = energy per source photon in million electron volts.

For a single material, t would be the thickness traversed (in centimeters), and μ would be the total attenuation coefficient (cm^{-1}). For many layers, μt would be the sum of these factors for each layer. The total attenuation coefficients, dose factors f , and buildup factors for gamma rays passing through lead are given in Table I. Thus, for example, the dose at 70 cm from a 100-curie ($100 \times 3.7 \times 10^{10} \text{ sec}^{-1}$) source of 1.0-MeV gamma rays in a lead box shield (strictly, a spherical shell container) of wall thickness 10 cm is

$$D(70, 10, 1.0) =$$

$$\frac{100 \times 3.7 \times 10^{10} \cdot 3.19 \cdot e^{-7.73} \cdot 1.73 \times 10^{-6} \cdot 1.0}{4\pi \cdot (70)^2} \quad \text{rad/hr}$$

$$= 1.46 \times 10^{-1} \text{ rad/hr}$$

$$= 146 \text{ mrad/hr}$$

Neutron attenuation is more difficult to calculate because of the much less regular variation of cross

sections with energy and the greater variety of possible interactions. Simple estimates of the dose rates outside thick concrete shields can be obtained by use of the formula

$$D(R, t, E) = \frac{S e^{-\Sigma_r t} F}{4\pi R^2}$$

where

Σ_r = macroscopic effective removal cross section of the concrete used, calculated from the data of Fig. 1 (cm^{-1}); for ordinary concrete, $\Sigma_r \approx 0.086 \text{ cm}^{-1}$;

F = dose rate, rem/hr, per unit flux at the source energy; for a fission source and a thick shield, F is the same as that for neutrons at about 8 MeV; roughly, for $1 < E < 10 \text{ MeV}$, $F \approx 1/7000 \text{ rem/hr per neutron/cm}^2 \text{ sec}$;

S = source strength, neutrons/sec.

EVERITT P. BLIZARD

References

1. Cranberg, L., *et al.*, *Phys. Rev.*, **103**, 662 (1956).
2. Keepin, G. R., Wimett, T. F., and Zeigler, R. K., *Phys. Rev.*, **107**, 1044 (1957).

TABLE I. DATA FOR GAMMA-RAY ATTENUATION IN LEAD

Photon Energy, E (MeV)	Attenuation Coefficient, μ (cm^{-1})	Dose Factor, f , (rad/hr per MeV/ cm^2 sec)	Buildup Factor, $B(\mu t)$, for a μt of						
			1	2	4	7	10	15	20
0.50	1.64	1.84×10^{-6}	1.24	1.42	1.69	2.00	2.27	2.65	2.73
1.0	0.773	1.73×10^{-6}	1.37	1.69	2.26	3.02	3.74	4.81	5.86
2.0	0.516	1.47×10^{-6}	1.39	1.76	2.51	3.66	4.84	6.87	9.00
4.0	0.475	1.19×10^{-6}	1.27	1.56	2.25	3.61	5.44	9.80	16.3
6.0	0.492	1.06×10^{-6}	1.18	1.40	1.97	3.34	5.69	13.8	32.7
10.0	0.552	0.94×10^{-6}	1.11	1.23	1.58	2.52	4.34	12.5	39.2

3. Goldstein, H., "Fundamental Aspects of Reactor Shielding," Reading, Mass., Addison-Wesley Press, 1959.
4. Way, K., and Wigner, E. P., *Phys. Rev.*, 73, 1318 (1948).

Cross-references: CROSS SECTIONS AND STOPPING POWER, COMPTON EFFECT, ELECTRON, FISSION, FUSION, NEUTRON, NUCLEAR REACTIONS, NUCLEAR REACTORS, NUCLEONICS, PHOTOELECTRICITY, POSITRON.

RECTIFIERS

Rectification (Electric). Energy is most conveniently transmitted and distributed by means of galvanic electric currents in metallic conductors (see CONDUCTIVITY, ELECTRICAL). In this form energy is easily transported with very low losses, easily accessible, and easy to control (see ELECTRIC POWER GENERATION). Transmission by electric current may also communicate intelligence, such as images, electroacoustic signals or counting pulses. Most human activities depend on some form of electric energy transport or distribution, such as power, telephone, radio, television, and most control functions.

Generators of electric energy from other physical forms normally utilize a mechanical intermediary (involving rotation or harmonic oscillation) followed by an electromagnetic generator. The output current is then equally alternating in both directions, it can thus be transformed to any desired current and voltage level. Alternating current is ideal for generating and transporting energy alone because no net electronic or electrolytic charge or material transport is required. On the other hand, in usage, whenever electrical energy is stored in batteries, energizes vacuum tube amplifiers, or is used for electrochemical separation or particle acceleration, the permanent and irreversible transport of charges is mandatory, hence a direct current is needed.

Rectifiers provide the physical means which achieve electric rectification, comprising all the elements which connect a complete ac circuit to a complete dc circuit, without being part of either (see Fig. 3). One rectifier may consist of a plurality of rectifier diodes, their mode of interconnection is called a *rectifier circuit*.

Rectifier Diodes. These are unilaterally conducting component devices with two terminals, similar to resistors because they are passive (i.e., not generating electric energy) and nonreactive (i.e., not able to store electric energy). The difference is in their being essentially nonlinear (whereas resistors are linear), their differential resistance varying over a very wide range, depending only on the direction and magnitude of the current through the device, i.e., they are not time-dependent (see Fig. 1).

Rectifier Switching. Rectification implies the concept of switching, i.e., introducing a circuit element which has a resistance which varies instantaneously over such a wide range that it may (mathematically) be considered as a discontinuity.

Example: Semiconductor diode (see Fig. 4)

Circuit impedance: 20 ohm

Diode forward resistance: 0.028 ohm 1/700 of circuit

Diode reverse resistance hot: 500,000 ohm - 25,000 times circuit.

Controlled Rectifiers (Fig. 2). Controlled rectifiers are similar to rectifier diodes, except that they have two states of forward conductivity: forward blocking (same resistance characteristic as in the reverse direction, which is the same as in a diode) and forward conducting (same as in a diode). A control element (gate, grid) allows switching from the forward blocking to the forward conducting state. This control is very rapid and requires very little energy. Hence, the rectifier output can be controlled with a high amplification and with high speed. Gas tube controlled rectifiers are thyatrons. Semiconductor controlled rectifiers (SCR) also known as thyristors.

Electronic Rectifier. Vacuum Tube (Diode, Triode, etc.). Electronic rectifiers are based on the Edison effect (see ELECTRON TUBES, ELECTRONICS, THERMOIONICS). Thermally emitted electrons from a hot metal oxide cathode are propelled across a short gap in high vacuum to a cold metallic anode. The high velocity of thermal electrons and the absence of a gas, generating positive ions, assure ideal conditions for electric rectification, i.e., a flow of pure electrons. Because of their high speed, the high-frequency response is very good. The potential field of the driving voltage accelerates the electrons, hence the rectifier is

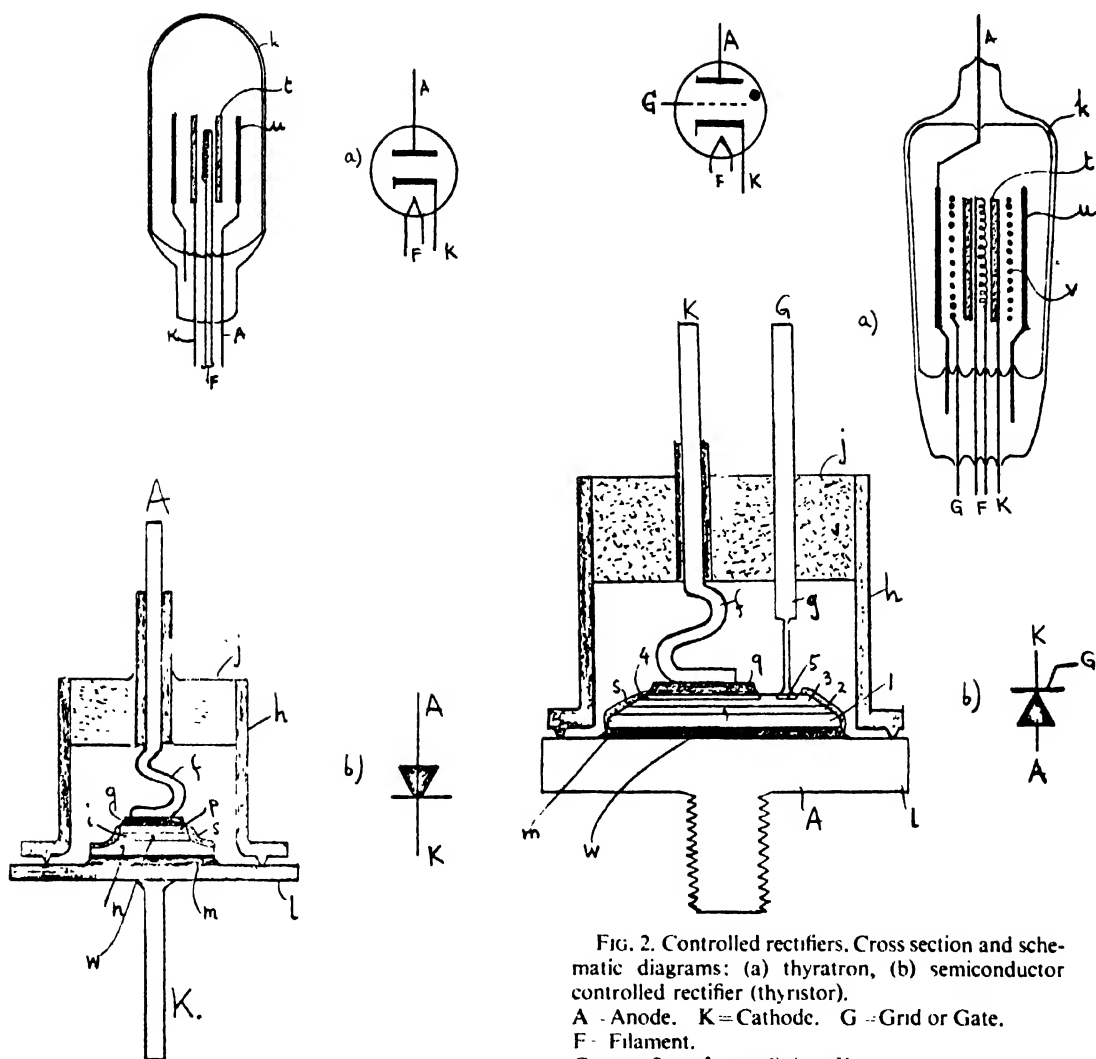


FIG. 1. Rectifier diode cross section and schematic diagram: (a) vacuum tube, (b) semiconductor.

A = Anode. K = Cathode. F = Filament.

Current flow (forward) A to K.

Voltage blocking (reverse) K positive, A negative.

f = Metallic anode lead, hermetically sealed.

h = Metallic case, welded to l, gas filled.

i = Intrinsic silicon (not doped with impurities).

j = Ceramic insulator.

k = Glass tube, evacuated.

l = Metallic base plate.

m = Metallic bond to cathode.

n = Negatively doped silicon.

p = Positively doped silicon.

q = Metallic bond to anode.

s = Surface of wafer, insulated.

t = Tubular cathode insulator coated with metal oxide.

u = Tubular outer electrode.

w = Wafer of silicon.

FIG. 2. Controlled rectifiers. Cross section and schematic diagrams: (a) thyatron, (b) semiconductor controlled rectifier (thyristor).

A = Anode. K = Cathode. G = Grid or Gate.

F = Filament.

Current flow (forward) A to K.

Reverse voltage blocking, K positive, A negative.

1 = p-type anode layer.

2 = n-type intermediary layer.

3 = p-type gate layer.

4 = n-type cathode layer.

5 = p-type contact for gate.

f = Metallic cathode lead, hermetically sealed.

g = Metallic gate lead.

h = Metallic case, welded to l, gas-filled.

j = ceramic insulator.

k = Glass tube, gas-filled.

l = Metallic base plate.

m = Metallic bond to anode.

q = Metallic bond to cathode.

s = Surface of wafer, insulated.

t = Tubular cathode insulator, coated with metal oxide.

u = Tubular outer electrode.

v = Helicoidal grid structure.

w = Wafer of silicon.

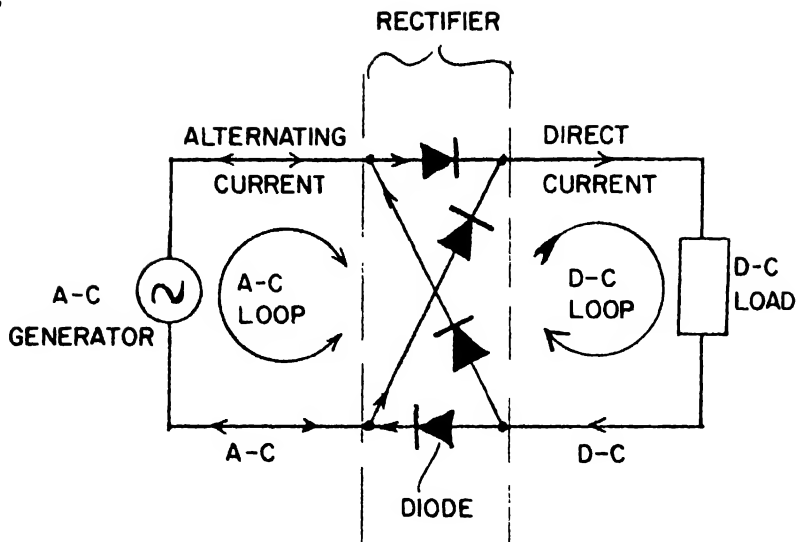


FIG. 3. Rectifier circuit: single-phase bridge. Four rectifier diodes are connected together to achieve full rectification of both (negative and positive) waves of a single phase alternating current. The alternating current flowing in the closed loop (ac circuit) on the left-hand side is converted into a direct current flowing in the closed loop (dc circuit) on the right-hand side.

essentially a linear resistor in the forward direction and a good insulator in the reverse direction. The applications of the vacuum tube are far beyond the scope of simple rectification. All the grid- and beam-controlled tubes are also rectifier diodes, although generally not defined as such. This is because the resistance changes (see MODULATION) in the current-carrying direction are technically and economically much more important.

Thyratron. Rectifier tubes containing a low-pressure gas instead of a vacuum have the property that the resistance in the current-carrying stage is very low but not linear, as in a vacuum tube. This makes the device useless for variable resistance control (amplification), but much more efficient for rectification; a major resistance in the current flow direction is eliminated. This reduction of resistance is caused by impact ionization of the gas, hence the thyratron responds less rapidly than the vacuum tube. The ionized gas is usually a pure element (e.g., mercury, hydrogen, krypton) to avoid chemical reactions between ions and electrodes. Current is carried both as positive ions and negative electrons. Thus, reversing the voltage on the gap results in a definite period of current reversal (ion sweep-out). This severely limits the useful operating frequency. A control grid allows one to select the time at which the thyratron becomes conductive ("firing"). Once conduction is initiated, the grid field is neutralized by the ionized gas, hence grid control has no effect (see Controlled Rectifier above).

Excitron (Mercury-arc Rectifier). This is similar to the thyratron, except the cathode is a pool of liquid mercury. An arc is sustained between the cathode and an exciting anode; conduction to the

main anode is controlled both by anode voltage and a control grid.

Ignitron. Replacing the heated solid cathode of a diode by a pool of liquid mercury gives an ignitron. To initiate conduction, a localized spot on the liquid surface (cathode spot) is forcibly overheated (ignited). Selecting the firing time (controlled rectifier) is achieved by energizing an ignitor which creates a cathode spot at a definite time. Ignitors are silicon carbide rods dipping into the mercury; subjected to a brief current pulse, a surface arc occurs at the mercury-silicon carbide interface. The liquid is locally driven to a very high temperature by forced electron emission and ion impact, resulting in violent evaporation and ionization of the mercury. Once conducting, the ignitron has the properties of a gas-filled tube (thyratron); it also requires the same deionization time and ion sweep-out current. Excess mercury vapor precipitates on the cold walls of the tube, flowing back into the liquid cathode. The ignitron is particularly applicable to very high power. It presents the advantages of high efficiency, high reliability and small size.

Semiconductor Rectifier. Crystallized semimetals (such as selenium, copper oxide, germanium or silicon) and some organic compositions can be used to make devices which rectify electric currents (see SEMICONDUCTORS and TRANSISTOR). Semiconductors carry current by excess electrons or electron vacancies (carriers) moving in the solid crystal lattice. The transfer of charges is very rapid and driven by very small potential differences. The polarity of a semiconductor is not determined by the material itself but by relatively few impurity nuclei ("doping") substituted in the crystal. Impurities (compared to the base material) have either an excess (*n*-type,

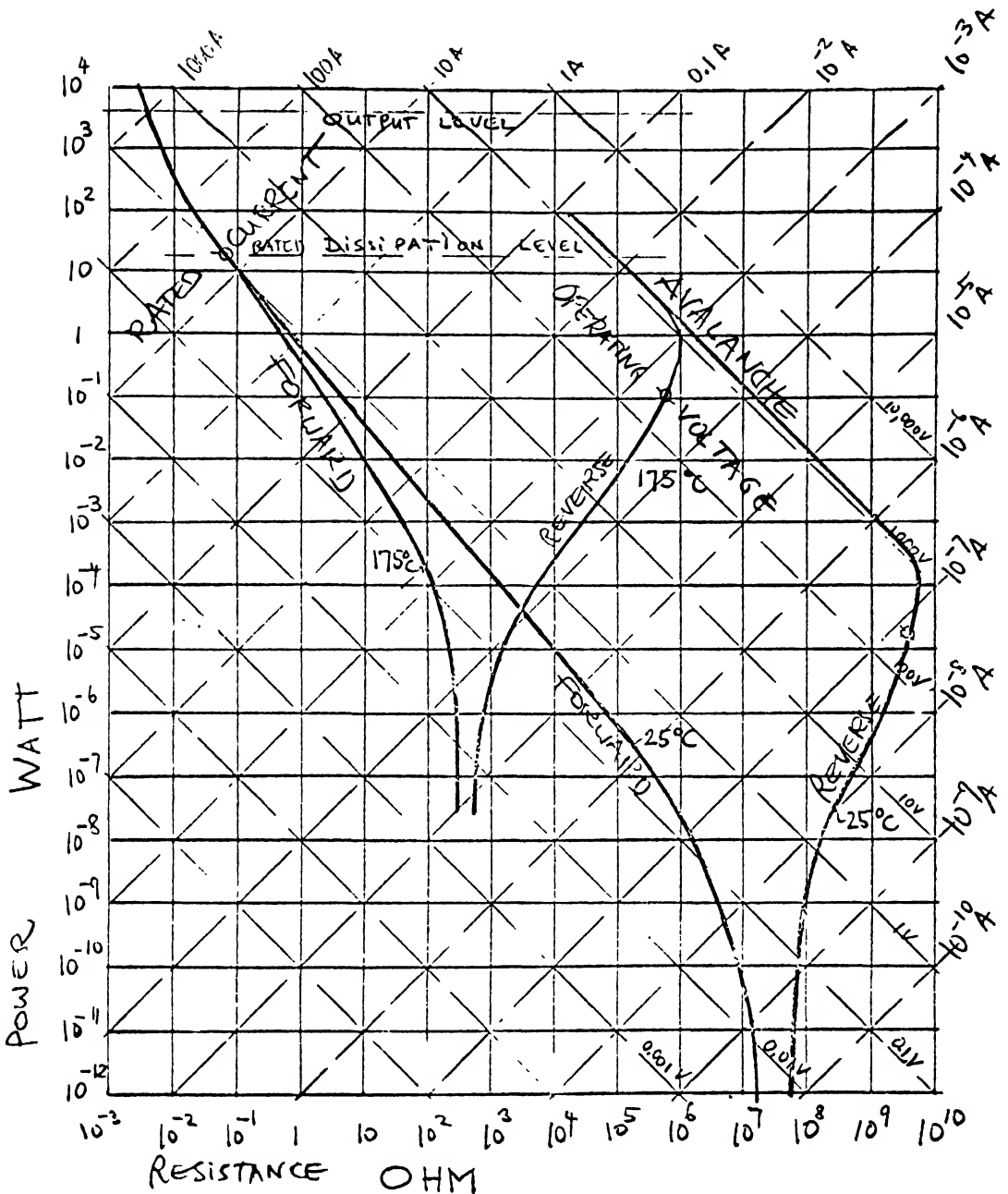


FIG. 4. Semiconductor rectifier diode, forward and reverse characteristics at 25 and 175 C. Power vs resistance diagrams, shown in the same quadrant. Influence of temperature at very low power levels only. Logarithmic scales. Current and voltage scales 45° angles. Constant resistance lines are vertical. Diode characteristics in forward and reverse avalanche direction are essentially in the direction of constant voltage, whereas in reverse blocking they are in the direction of constant current.

negative, conducting by "electrons") or a deficiency (*p*-type, positive, conducting by "holes") of nuclear charges and hence electron shells. In the rigid lattice, the nuclei and normal electron shells are immobile. Excess or defect carriers are freely mobile. The density of these majority

carriers is determined by the relative content of impurities in the crystal. The background of the lattice with its mutually neutralizing nuclei and electron shells does not contribute to the conduction or to the distribution of potentials.

Semiconductors conduct both by majority

carriers (e.g., holes in *p*-type), and minority carriers (e.g., electrons in *p*-type) if such are injected by a junction with the opposite polarity material, e.g., majority carrier electrons coming from *n*-type material injected into *p*-type (see TRANSISTOR). Injected minority carriers are ultimately trapped and recombined with majority carriers; however, they transfer a major quantity of charge from one zone to another, depending on the "lifetime" (see SEMICONDUCTOR) of these minority carriers.

Semiconductor Rectifier Diodes (Fig. 1). They contain one thin, flat wafer consisting of a single crystal (e.g., silicon). The wafer is brazed to metallic electrodes (anode and cathode) on its two opposing flat sides. The rim of the wafer is insulated (by oxidation, insulating resin or fused glass). Within the wafer, a junction is established by heavy doping with *p*-type impurities on its anode face and *n*-type impurities on the cathode face. The junction consists of an intermediate, thin, flat zone in which the density of both *p*-type and *n*-type dopants is very low.

Applying a forward bias (P-positive, N-negative) to the device injects majority carriers from both zones through the junction into the opposite zone. Attracted by the opposite potential, they effect a total transfer of available charges; i.e., a current flows with only a low driving potential difference. An unlimited number of carriers can flow across the small potential (energy level) barrier between the zones, carriers are replenished by metal-to-semiconductor brazed joints on both faces of the wafer.

Reversing the bias at the junction (N-positive, P-negative) reverses the flow of carriers. Majority carriers from both zones are displaced away from the junction, a potential wall is created by depleting the crystal of its mobile carriers. The immovable charges of the lattice-bound nuclei which are now uncompensated by the displaced mobile charges, create a high potential wall. A very small reverse current flows, sustained only by the thermally generated minority carriers of both zones which are swept across the junction. When applying a very high potential (e.g., 1000 volts), these highly accelerated carriers generate more carriers by avalanche multiplication due to impact with the lattice; above a certain voltage, so many carriers are generated that the reverse characteristic remains at a constant voltage at any current level. As Fig. 4 shows, semiconductor diodes have a very low forward voltage drop (e.g., 1 volt) and the forward resistance is not constant but decreases with increasing current. The reverse current is negligibly small, except at very high voltage where the reverse resistance is negligible.

Varying the semiconductor material (mainly germanium and silicon), the impurity content, the distribution of impurities across the junction, the area of the junction, and its peripheral configuration (planar, mesa, cut wafer) allows for a multitude of possible designs, each preferred for certain applications. Silicon wafers with diffused impurities (e.g., boron and phosphorus) are

used for high-power, low-frequency rectifier diodes. Planar junctions on the surface of a solid wafer, in which the impurities are diffused under a layer of protective oxide, give diodes with good high-frequency and low-noise response. Junctions are also made by recrystallizing silicon from an alloy melt (alloyed junctions) or by epitaxially depositing pure silicon, with measured impurities, from the vapor phase upon a solid wafer.

Semiconductor Controlled Rectifiers (Fig. 2). Similar to transistors (rather than diodes) they consist of four layers of semiconductor material forming three closely adjacent junctions (against two in a transistor). Only three layers are connected to outside electrodes. Figure 2 is an example: Reverse bias (e.g., anode negative) blocks the flow of current similar to a diode between anode layer 1 and *n*-type intermediary layer 2. Forward bias (e.g., anode positive) blocks the flow of current similar to a diode between *p*-type gate layer 3 and *n*-type layer intermediary 2. Initiating a current between gate 5 and cathode 4 injects electrons into the gate layer 3; these are attracted into the adjacent reverse biased junction (3-2) where they attract and multiply due to (forward current) carrier injection by the two outer junctions (1-2 and 3-4). Counter-flow of opposite carriers through the same junction results in a very low voltage drop.

Reverse and forward blocking characteristics are similar to diode reverse characteristics (see Fig. 4). Forward conducting characteristics are similar to diode forward characteristics.

Mechanical Rectifier. A coil of wire rotating in a magnetic field generates an alternating voltage; if the ends of many similar coils are connected to "commutator bars" (arranged on a cylindrical commutator), the output of two stationary brushes contacting the bars is a direct current. This dynamo is of historic and economic significance. Trial-and-error improvements have made this combination of rotating machine and mechanical rectifier an efficient and economic system. Mechanical rectifiers separated from the moving coils (i.e., a commutator driven by a synchronous motor) are workable but less reliable (because subject to wear) and less economic than other rectifiers.

Electrolytic Rectifier. Some electrolytic batteries can serve as rectifiers because of the rapid and thorough polarization of one of the electrodes, whereas the other electrode may carry the current rather easily. The system is permanently changed by the flow of current; hence, it is not reliable or economical. Its speed of response is low, depending on the slow diffusion of ions in the electrolyte.

EDWARD J. DIEBOLD

Cross-references: CONDUCTIVITY, ELECTRICAL; ELECTRIC POWER GENERATION; ELECTRON TUBES; ELECTRONICS; MODULATION; SEMICONDUCTORS; THERMIONICS; TRANSISTOR.

REFLECTION

If a perfectly smooth and flat interface exists between two homogeneous media, the ratio of the reflected to the incident intensity of light striking the interface is called the *reflectivity*. A real surface however, is not ideally smooth and undistorted so that the situation is complicated by surface conditions such as films, roughness, and disorder in the crystal lattice at the surface introduced by the polishing process. The measured ratio is thus usually termed the *reflectance* to distinguish it from the former more- or less-idealized situation.

The reflectivity of a material is intimately connected with its band structure.¹ In a single atom, the electrons surrounding the nucleus can have only certain discrete energies. In a solid or liquid, neighboring atoms interact with each other changing these discrete energy levels into energy bands. The energy band structure depends not only on the type of atoms involved but also on the interatomic spacing, and hence in a crystalline material on the lattice structure. In a noncrystalline solid or in a liquid, the correlation function takes the place of the crystalline lattice. The highest occupied energy band may be completely filled with electrons, as in the case of a dielectric or semiconductor, or only partially filled, as for a metal. If the band is not completely full, electrons may absorb energy from the incident light beam and be raised to higher energies in the band, thus affecting the reflectivity of the material. These *intraband* transitions of the conduction electrons, or, in the case of a nearly filled band, holes (the absence of electrons), largely determine the reflectivity of metals and some semiconductors in the infrared region of the spectrum. At shorter wavelengths, the light has sufficient energy to raise the electrons from one energy band to another, and the reflectivity is then largely determined by these *interband* transitions. At still shorter wavelengths, usually in the vacuum ultraviolet, the energy in the incident light beam can set the conduction electrons into collective oscillation so that a plasma resonance occurs in the reflectivity spectrum.

In dielectric materials, where the occupied band having the highest energy is completely filled and the gap between allowed energy bands is large, the energy in the incident light beam cannot be absorbed either in intraband transitions or, except at very short wavelengths, in interband transitions. Such materials are therefore often transparent over an extended wavelength region in the ultraviolet, visible and near infrared, and have a nonzero reflectivity in this wavelength region only because of the difference in the speed of light in the material and in the surrounding medium. However various types of imperfections, termed F centers, M centers, etc., may exist in the lattices of crystalline materials, and electrons trapped in such centers can strongly absorb certain energies, thus affecting the reflectivity. In the intermediate and far infrared, dielectrics and many semiconductors absorb because the

incident light excites vibrations in their crystal lattices. Since the frequencies of these lattice vibrations are quantized, this so-called phonon absorption results in maxima in the reflectivity spectra, termed *restrahlen* bands.

Other mechanisms also contribute to the reflectivity. Some which should be mentioned are transitions between a localized impurity level and an energy band, absorption by a bound hole-electron pair or exciton, and indirect or phonon-assisted interband transitions.

Electromagnetic Theory. The reflectivity of any material can be calculated from electromagnetic theory^{2,3} if the optical constants of the material are known. This theory, which is based entirely on Maxwell's equations, is phenomenological in that it does not attempt to explain why materials behave as they do, but rather sets forth relationships which exist between various properties of the material. In order to calculate the reflectance of a material from this theory, the two optical constants, n and k , must be known. The index of refraction n is equal to the ratio of the phase velocity of light in vacuum to the phase velocity in the material. The extinction coefficient k is equal to $\lambda/4\pi$ times the absorption coefficient of the material and is thus a measure of the fraction of light absorbed by a unit thickness of the material. These parameters, which are frequently combined into the complex refractive index $\tilde{n} = n - jk$, arise in the solution to the wave equation and completely describe all the optical properties of the material.

The equation for the propagation of an electromagnetic wave can be obtained directly from Maxwell's equations, and in Gaussian units can be written

$$\nabla^2 \mathbf{E} = \frac{\epsilon}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} + \frac{4\pi\sigma}{c^2} \frac{\partial \mathbf{E}}{\partial t} \quad (1)$$

where ∇^2 is the Laplacian operator $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$, \mathbf{E} the electric field strength of the traveling

wave, t the time, c the velocity of light and ϵ and σ the dielectric constant and conductivity of the material, respectively, at the frequency of the wave. The solution representing a plane wave traveling in the z direction is

$$E = E_0 e^{j\omega(t - nz/c)} \quad (2)$$

where $\omega = 2\pi\nu$ is the angular frequency of the wave, E_0 the amplitude, and \tilde{n} the complex refractive index. By matching E_0 and its first derivative on either side of a smooth, plane interface between a transparent material of index n_0 , the medium of incidence, and a second material of index \tilde{n}_1 , one can obtain r , the ratio of the reflected to the incident amplitude, called the *amplitude reflection coefficient*:

$$r = \frac{\eta_0 - \tilde{\eta}_1}{\eta_0 + \tilde{\eta}_1} \quad (3)$$

where η_0 and $\bar{\eta}_1$ are the effective indices. At normal incidence, $\eta_0 = n_0$ and $\bar{\eta}_1 = \bar{n}_1 = n_1 - jk_1$.

At non-normal incidence, Snell's law may be used to determine the angle of refraction ϕ_r corresponding to a given angle of incidence ϕ_i . For the case when both materials are nonabsorbing, Snell's law states

$$\frac{\sin \phi_i}{\sin \phi_r} = \frac{n_1}{n_0} \quad (4)$$

It is also necessary to specify the state of polarization of the incident light. Since \mathbf{E} is a vector quantity, it is always possible to resolve it into two components, the so-called p and s components, polarized parallel to and perpendicular to the plane of incidence (the plane containing both the incident beam and the normal to the surface). There are then two effective indices, η_p and η_s , for each medium at a given angle of incidence. For nonabsorbing materials they are defined as

$$\eta_s = n \cos \phi \quad (5)$$

$$\eta_p = \frac{n}{\cos \phi} \quad (6)$$

where ϕ is the angle of incidence or refraction in the medium of index n . In the medium of incidence, ϕ becomes ϕ_i while in the second medium, ϕ becomes ϕ_r . If the second medium is absorbing ($k \neq 0$), $\cos \phi$ becomes complex, making η_s and η_p also complex. They may then be most easily calculated from the following expressions:⁴

$$\bar{\eta}_1 = A - jB \quad (7)$$

$$\eta_p = C - jD \quad (8)$$

where

$$A^2 - B^2 = n_1^2 - k_1^2 = n_0^2 \sin^2 \phi_i \quad (9)$$

$$AB = n_1 k_1 \quad (10)$$

$$C = A \left[1 + \frac{n_0^2 \sin^2 \phi_i}{A^2 + B^2} \right] \quad (11)$$

$$D = B \left[1 - \frac{n_0^2 \sin^2 \phi_i}{A^2 + B^2} \right] \quad (12)$$

The *intensity reflection coefficient* or *reflectivity* R , defined as the ratio of the intensities of the reflected and incident light, is obtained by multiplying the amplitude reflection coefficient of Eq. (3) by its complex conjugate. At normal incidence for an absorbing material

$$R = \frac{(n_1 - n_0)^2 + k_1^2}{(n_1 + n_0)^2 + k_1^2} \quad (13)$$

At non-normal incidence, the expressions for R_s and R_p for absorbing materials become quite complicated. However, for nonabsorbing materials

$$R_s = \frac{\sin^2(\phi_i - \phi_r)}{\sin^2(\phi_i + \phi_r)} \quad (14)$$

and

$$R_p = \frac{\tan^2(\phi_i - \phi_r)}{\tan^2(\phi_i + \phi_r)} \quad (15)$$

Equations (14) and (15) are the intensity expressions for Fresnel's equations. From Eq. (15) it is seen that when $\phi_i + \phi_r = 90^\circ$, $R_p = 0$ so that all of the reflected light is polarized in the s direction. The angle of incidence for this case is called the polarizing angle and is given by

$$\tan \phi_i = \frac{n_1}{n_0} \quad (16)$$

A plot of the reflectivity R_p and R_s as a function of angle of incidence is shown in Fig. 1(a) for a transparent material in air and in Fig. 1(b) for an absorbing material. Note that the R_p curve for the transparent material goes to zero while the R_p curve for the absorbing material does not.

The phenomenon of total internal reflection is important for transparent materials. If light passes from a more optically dense (higher index) medium of index n_0 to a less optically dense (lower index) medium of index n_1 , when $\sin \phi_i = n_1/n_0$, the angle of refraction in the less dense medium is 90° , as can be seen from Eq. (4). At larger angles of incidence, the light is totally internally reflected in the more dense medium. This method of obtaining a surface whose reflectivity is 100 per cent is widely used in optical instruments such as binoculars and periscopes, which contain internally reflecting prisms.

Whenever light is reflected from a surface, a change in phase occurs in the electric vector. For the case of a normal incidence reflection in air from a dielectric, the phase change β is 180° , while for a reflection in the more dense medium, β is 0° . When the reflection is from an absorbing material, β depends on the optical constants of the material,⁵ and $\tan \beta$ is given by the ratio of the imaginary to the real part of Eq. (3). If the medium of incidence is air,

$$\tan \beta = \frac{2k}{n^2 - k^2 - 1} \quad (17)$$

for a normal incidence reflection from an absorbing material with $\bar{n} = n - jk$. At non-normal incidence, the phase change on reflection is different for the p and s components. By a technique called ellipsometry,⁶ this difference in phase can be measured and used, along with other data, to determine the optical constants of a material. If a surface film is present, the phase difference between the p and s components can also be used to measure its growth. Films as thin as a monolayer or less can be detected in this way.

As was indicated at the beginning of this section, Maxwell's equations by themselves do not give sufficient information to relate the optical constants of a material to basic atomic parameters. A fundamental problem in the theory of the optical properties of solids is thus to supplement electromagnetic theory in such a way as to relate the optical constants of a material to nonoptical quantities which can be experimentally determined. If Eq. (2) is substituted into Eq. (1), one obtains the basic

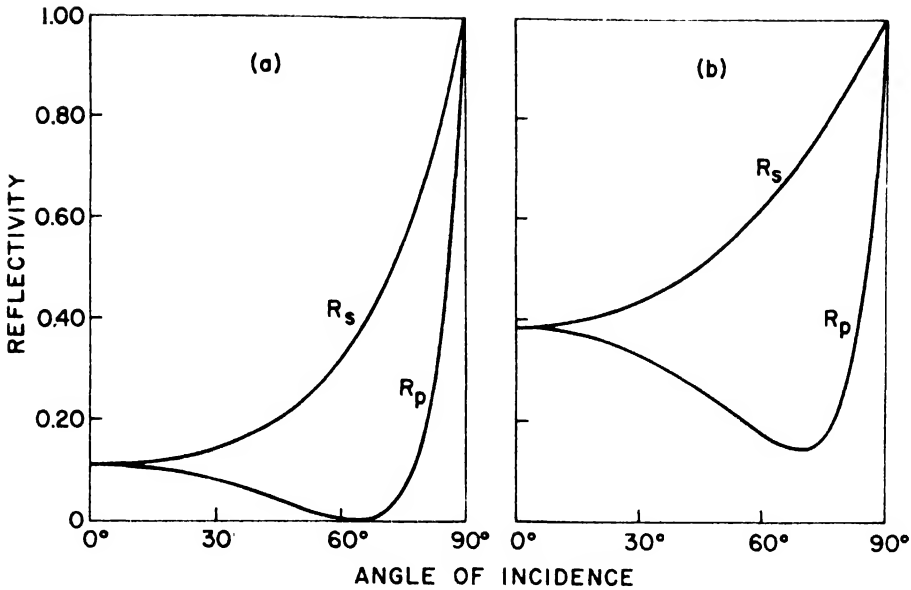


FIG. 1. Graph of the reflectivity R_p and R_s as a function of angle of incidence for (a) a transparent material ($n = 2$) in air, and (b) an absorbing material ($n = 2$ and $k = 2$) in air.

equations relating the optical constants to the dielectric constant and conductivity measured at optical frequencies:

$$n^2 - k^2 = \epsilon \quad (18)$$

$$nk = \sigma/\nu \quad (19)$$

The problem is how to relate ϵ and σ to the dielectric constant and conductivity at zero frequency, ϵ_0 and σ_0 , or to atomic parameters such as the effective mass m^* , number of free carriers per cubic centimeter N , oscillator strength f , etc. A classical simple harmonic oscillator theory first proposed by H. A. Lorentz in 1880 has been successfully used to solve this problem for materials in which the absorption is caused by forced vibrations, while a companion classical theory proposed by P. Drude in 1901 has had similar success for materials where the absorption is due to free electrons and holes. Quantum mechanical treatments give rise to equations having the same form as the classical ones, but containing parameters whose values can be calculated rather than having to be empirically determined. Although these theories will not be discussed here, it can be pointed out that, to a very good approximation, the reflectivity of good conductors in the infrared region is given by⁷

$$R \approx 1 - \left[\frac{2\omega}{\pi\sigma_0} \right]^2 [(1 + \omega^2\tau^2)^{-1} - \omega\tau] \quad (20)$$

where the relaxation time τ is given by

$$\tau = \frac{m^*\sigma_0}{Ne^2} \quad (21)$$

and e is the electronic charge. At sufficiently long wavelengths, $\omega\tau \ll 1$ and the reflectivity is then given by the Hagen-Rubens relation

$$R = 1 - \left[\frac{2\omega}{\pi\sigma_0} \right]^2 \quad (22)$$

which may also be obtained directly from Eqs. (13), (18) and (19). Both Eqs. (20) and (22) are in good agreement with experiment.

Reflectance of Real Surfaces. In the preceding discussion, it has been assumed that there are two homogeneous media separated by a perfectly smooth and flat interface. In actual fact this situation does not occur, and it is necessary to consider how the deviation of real surfaces from this model affects the observed reflectance. Consider first the effect of surface roughness. For a perfectly smooth surface, all light is reflected at the specular angle, which is equal to the angle of incidence. At the other extreme, the angular dependence of light reflected from a perfectly diffuse reflector is independent of the angle of incidence, and Lambert's law then holds. This law may be stated

$$I_r(\theta) = I_0 \cos \theta \quad (23)$$

where I_0 is the total amount of light reflected per unit area normal to a perfectly diffuse plane reflector, and $I_r(\theta)$ that reflected per unit area at an angle θ to the normal. The reflectance of actual rough surfaces can be separated into "specular" and "diffuse" components, although the diffuse component usually does not obey Lambert's law exactly. If the heights of the surface irregularities are of the order of a wavelength or more, the reflectance is mostly diffuse

and depends strongly on the shape of the irregularities. However, if they are very small relative to the wavelength, most of the light is reflected specularly, the amount depending only on the height of the irregularities. For a random height distribution of surface irregularities, the specular reflectance at normal incidence R_s is given by⁸

$$R_s = R_0 e^{-(4\pi h/\lambda)^2} \quad (24)$$

where R_0 is the total reflectance of the surface, h the rms height of the surface irregularities, and λ the wavelength. R_s/R_0 is very sensitive to small values of h/λ . For example, if h is as small as 0.025λ (in the visible h would then be only one-half a microinch), the specular reflectance will be decreased by 10 per cent. The surfaces of mirrors used in optical instruments must thus be extremely smooth. Typically the rms roughness of such mirrors is less than 0.1 microinch and may be as small as 0.01 microinch.

The reflectance may also be affected by distortion of the crystal lattice at the surface caused by the polishing process. The amplitude penetration depth δ of light into an absorbing medium, given by

$$\delta = \lambda/2\pi k \quad (25)$$

is usually only a few hundred angstroms or roughly one microinch for metals or semiconductors in the intrinsic absorption region. On the other hand, lattice distortion introduced by optical polishing may extend for many thousand angstroms below the surface, so that the reflection takes place entirely in the disturbed surface layer. Fortunately, the lattice distortion on the surface can be nearly eliminated in many cases by using proper electropolishing techniques.⁹ Optical polishing can also produce changes in the reflectance of dielectrics and non-crystalline materials such as optical glass.

Changes in the energy band structure at the surface may also be caused by the crystal structure of evaporated films and by the presence of surface states in semiconductors. Both of these effects may influence the reflectance.

Finally, surface films can have a large effect on the reflectance. Although naturally occurring oxide films are important mainly in the ultraviolet, the reflectance may be substantially modified at any wavelength by overcoating the material with an evaporated thin film. Over a limited wavelength region, nearly any desired reflectance characteristic may be obtained in this way, and this technique is widely used in the optical industry.¹⁰ Perhaps the most familiar example of the application of a thin film is the antireflection coating on lenses. Others are the "cold" mirrors used in projection systems, the multilayer coatings which control the temperature of space vehicles by adjusting the reflectance of their outer surfaces, and finally the highly reflecting, low-absorbance dielectric multilayer films used with lasers.

H. E. BENNETT

References

1. Callaway, J., "Energy Band Theory," New York, Academic Press, 1964; Bube, R. H., "Photoconductivity of Solids," New York, John Wiley and Sons, Inc., 1960.
2. Born, M., and Wolf, E., "Principles of Optics," New York, Pergamon Press, 1959; Ditchburn, R. W., "Light," Second edition, London, Blackie and Son Ltd., 1963.
3. Stratton, J. A., "Electromagnetic Theory," New York, McGraw-Hill Book Co., 1941; Slater, J. C., and Frank, N. H., "Electromagnetism," New York, McGraw-Hill Book Co., 1947.
4. Abelès, F., "Progress in Optics," Vol. 2, Amsterdam, North-Holland Publishing Co., 1963.
5. Bennett, J. M., *J. Opt. Soc. Am.*, **54**, 612 (1964).
6. McCrackin, F. L., Passaglia, E., Stromberg, R. R., and Steinberg, H. L., *J. Res. Natl. Bur. St.*, **67A**, 363 (1963).
7. Seitz, F., "The Modern Theory of Solids," New York, McGraw-Hill Book Co., 1940; Bennett, H. E., Silver, M., and Ashley, E. J., *J. Opt. Soc. Am.*, **53**, 1089 (1963).
8. Bennett, H. E., and Porteus, J. O., *J. Opt. Soc. Am.*, **51**, 123 (1961); Porteus, J. O., *J. Opt. Soc. Am.*, **53**, 1394 (1963); Beckmann, P., and Spizzichino, A., "The Scattering of Electromagnetic Waves from Rough Surfaces," New York, The Macmillan Co., 1963.
9. Donovan, T. M., Ashley, E. J., and Bennett, H. E., *J. Opt. Soc. Am.*, **53**, 1403 (1963); Holland, L., "The Properties of Glass Surfaces," London, Chapman and Hall, 1964.
10. Heavens, O. S., *Rep. Progr. Phys.*, **23**, 1 (1960); Heavens, O. S., "Optical Properties of Thin Solid Films," London, Butterworths Scientific Publications, 1955.

Cross-references: ENERGY LEVELS; OPTICS, GEOMETRICAL; OPTICS, PHYSICAL; POLARIZED LIGHT; REFRACTION; SEMI-CONDUCTORS.

REFRACTION

Refraction is the name given to the bending of a ray of light as it crosses the boundary separating two transparent media having differing propagation velocities.

In an attempt to discover a law connecting the directions of the light rays in the two media, W. Snell (1621) observed that there is a constant ratio between the lengths PB, PC of the two rays (Fig. 1) measured from the point of incidence P to any line such as DD' drawn parallel to the normal PN at the point of incidence. Later Descartes recognized that Snell's construction is equivalent to the mathematical expression

$$n \sin I = n' \sin I'$$

when I, I' are the angles of incidence between the two rays and the normal, and n, n' are the refractive indices of the two media respectively. The ratio of n' to n is called the "relative" refractive index of the two media.

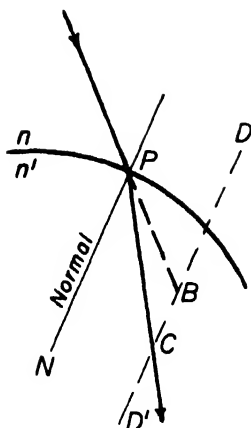


FIG. 1. Illustrates Snell's construction.

Huygens' Wavelets. C. Huygens (1690) attempted to explain how a wave front progresses through a transparent medium. He supposed that each point on the wave front acts as an independent source of wavelets which expand at the velocity of light; the new wave front being the common envelope of all the little wavelets. After an instant of time, each point in the new wave front becomes a source of new wavelets, and so on. In a homogeneous isotropic medium the wavelets will be spheres, and the new wave front is a parallel curve to the original wave front. The light energy travels along *rays* which are everywhere perpendicular to the wave fronts. Malus (1808) showed that the rays and wave fronts remain always orthogonal as the beam of light from an object point traverses an optical system.

The absolute refractive index of a medium is defined as the ratio of the velocity of light in vacuum to its velocity in the medium. Because the presence of matter has the effect of making light travel more slowly, all absolute refractive indices are necessarily positive and greater than unity. Since the frequency of the light waves must remain constant, the wavelength in a dense medium will be less than the wavelength in vacuum in proportion to the refractive index of the medium. Refractive indices of liquids and crystals generally drop with increasing temperature; glasses are, however, often exceptional in this regard.¹

Refractive indices range from as low as 1.3 for some liquids such as water, through 1.5 to 1.9 for various types of glass, 2.5 for diamond up to as high as 4.0 for some materials (germanium) in the infrared. These exceptionally high indices are generally associated with complete opacity in the visible part of the spectrum.

Dispersion. It is found that the refractive index of all common materials rises with increasing frequency of the light (shorter wavelengths) leading to the phenomenon of *dispersion*. Because of this, a glass prism bends blue light more than red, thus spreading a ray of white light into its

component colors (Fig. 2). For a prism of very small angle A , the deviation angle D is given by

$$D = A(n - 1)$$

and the angular dispersion Δ between wavelengths a and b is given by

$$\Delta = A(n_b - n_a)$$

It has been customary to define the dispersive power w of a material as the ratio of the dispersion between the F and C Fraunhofer lines to the mean deviation, i.e., the deviation for the D Fraunhofer line. Thus

$$w = (n_F - n_C)/(n_D - 1)$$

The vacuum wavelengths of these lines are $C = 0.6563\mu$, $D = 0.5893\mu$, and $F = 0.4861\mu$. In the optical industry, the reciprocal of the dispersive power is, however, more generally used; i.e.,

$$V = 1/w = (n_D - 1)/(n_F - n_C)$$

This so-called " V -value" or "*Abbe number*" of optical glass ranges from about 25 for the densest flints up to about 70 for the lightest fluor crowns. Liquids and crystals are known in which the V -numbers range from about 16 (methylene iodide) to 95 (calcium fluoride).

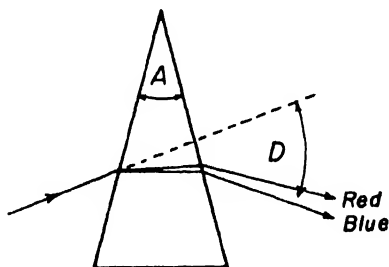


FIG. 2. The dispersion caused by a prism.

Many attempts have been made to develop a formula connecting the refractive index of a material with the wavelength of the light. The best known and most comprehensive is that proposed by Sellmeier,² namely,

$$n^2 = a + \sum \frac{b}{c^2 - \lambda^2}$$

Here n is the refractive index corresponding to wavelength λ , while a , b , and c are constants. The constant c represents the center of an absorption band for which the refractive index becomes infinite; there are thus as many terms under the summation sign as there are absorption bands to be considered. For most transparent materials, it is sufficient to include one absorption band in the ultraviolet and one in the infrared.

This general formula may be reduced to a simpler form for a limited spectral range, some well-known simplifications being

$$n = a + b/\lambda^2 + c/\lambda^4 + \dots \quad (\text{Cauchy})$$

$$n = 1 + b/(c - \lambda) \quad (\text{Hartmann})$$

$$n = a + b/\lambda + c/\lambda^{7/2} \quad (\text{Conrady})$$

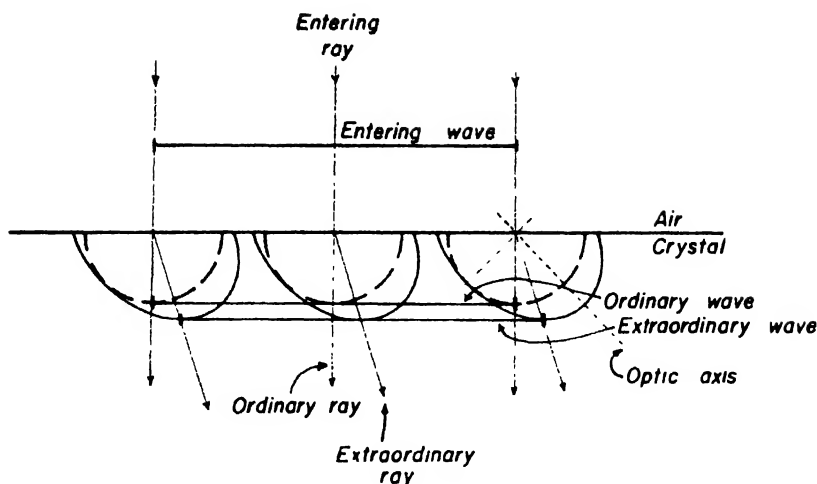


FIG. 3. Explanation of double refraction.

Recently Herzberger³ has shown that the refractive index of a transparent substance such as glass can be accurately represented by a four-constant formula of this type:

$$n = a + b\lambda^2 + cL + dL^2$$

where $L = 1/(\lambda^2 - 0.028)$.

Birefringence, or Double Refraction. Glasses, liquids, and some (e.g., cubic) crystals are *isotropic*, meaning that the velocity of light in the medium is independent of direction and the Huygenian wavelets are spherical. However, there are other crystals which are *anisotropic*. These have the property that a beam of light on entering the crystal is split into two perpendicularly polarized beams, one of which (the "ordinary") behaves normally, while the other (the "extraordinary") behaves abnormally in that the wave front is not perpendicular to the direction of propagation.

These two wave fronts can be explained by supposing that two Huygenian wavelets are formed at any point of incidence, the ordinary wavelet being spherical while the extraordinary wavelet is an ellipsoid, (Fig. 3).^{*} The advancing wave front is the common tangent to a row of wavelets, and the ray, or direction of travel, is found by joining the point of origin of the wavelet to the point of contact between the wavelet and the advancing wave front. The spherical wavelets thus yield a ray which is perpendicular to the wave front, but the ellipsoidal wavelets do not.

All anisotropic crystals possess either one or two directions ("optic axes") along which a ray of light is not divided into two. These are represented by the axes of symmetry of the ellipsoidal wavelets. Evidently if parallel light is travelling along an optic axis, the two advancing wave fronts will be parallel to one another and the two rays will coincide (see POLARIZED LIGHT).

^{*} R. W. Wood, "Physical Optics", 3rd ed., pp. 365-387, MacMillan, New York, 1943.

Refraction by an inhomogeneous Medium.

Suppose we have a parallel plate of material of which the refractive index varies laterally across the plate. A ray entering such a plate perpendicularly to its surface will suffer no refraction itself since the incidence angle is zero, but because there is a gradient of refractive index at right angles to the ray, the ray inside the plate will be bent towards the high-index region and will follow a curved path. This is readily seen if we remember that the ray will be perpendicular to the wave front and that each point on the wave front travels at a rate which is inversely proportional to the refractive index at that point. To determine the curvature of the ray inside the medium, we may refer to Fig. 4 in which CC' represents the surface of a nonhomogeneous medium, and $A'A$, $B'B$ represent two very close rays entering perpendicular to the surface. We shall suppose the velocity of light in the material to be v at A and $(v + dv)$ at B , the refractive indices being $(n + dn)$ and n respectively. Typical wavelets have been added at A and B , having radii AD and BE which are proportional to the velocities v and $(v + dv)$. Hence the refracted wave front is ED ; this intersects the surface at C , which is therefore the instantaneous center of curvature of the rays within the medium.

If r is the radius of ray AD ,

$$\frac{CA}{CB} = \frac{r}{r + dr} = \frac{AD}{BE} = \frac{v}{v + dv} = \frac{n + dn}{n}$$

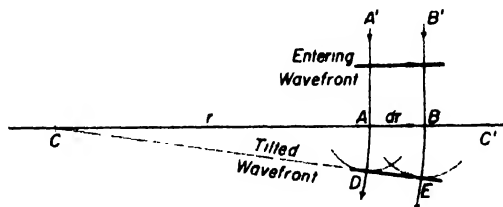


FIG. 4. Bending of a light ray in an inhomogeneous medium.

Neglecting second-order infinitesimals this gives

$$\frac{1}{r} = \frac{1}{n} \left(\frac{dn}{dr} \right)$$

The curvature of the ray is, therefore, proportional to the rate of change of refractive index in a direction perpendicular to the ray itself.

This property of inhomogeneous media provides the explanation of mirages and also the well-known atmospheric refraction which makes celestial objects appear to be raised up by as much as half a degree for objects situated near the horizon.

R. KINGSLAKE

References

1. Molby, F. A., *J. Opt. Soc. Am.*, **39**, 600-611 (1949).
2. Wood, R. W., "Physical Optics," Third edition, p. 470, New York, The Macmillan Co., 1943.
3. Heizerberger, M. J., *Optica Acta*, **6**, 197-215 (1959).

Cross-references: LIGHT; OPTICS, GEOMETRICAL; OPTICS, PHYSICAL; POLARIZED LIGHT; REFLECTION; WAVE MOTION.

REFRIGERATION

A refrigerator is an enclosure whose temperature is maintained substantially below the ambient temperature so that anything placed within may be kept cool. Early refrigerators, usually boxes immersed in cold running water or bathed in cold air from melting ice, were inconvenient and inefficient. Since the invention of vapor-compression and vapor-absorption refrigeration methods, mechanical refrigerators have become commonplace. More recently, the thermoelectric properties of semiconductors have been exploited and the magnetothermoelectric effects in semi-metals have been successfully employed in the laboratory especially at low temperatures.

A refrigerator may be considered a heat pump, for it extracts heat from a low-temperature region and delivers it to a high-temperature region, resembling a water pump which lifts water from a low-pressure region and delivers it to a high-pressure region. For household heating and air conditioning, such a system is so designed that, during the winter, heat is pumped from the earth or outside air into the building and, during the summer, heat is pumped in the reverse direction.

A refrigerator is rated by its *coefficient of performance*, which is defined as the ratio of the heat removed from the cold room per unit time to the net input power for operating the refrigerator; in symbols, $K \approx Q_1/P$. Vapor-absorption and thermoelectric refrigerators have lower coefficients of performance than vapor-compression refrigerators, but they have other

characteristics which are superior, such as quietness of operation or compactness.

Vapor-compression Refrigerator. This machine consists of a compressor, a condenser, a storage tank, a throttling valve, and an evaporator connected by suitable tubes with intake and outlet valves, as shown schematically in Fig. 1. The refrigerant is a liquid which partially vaporizes and cools as it passes through the throttling valve. Among the common refrigerants are ammonia, sulfur dioxide, and certain halide compounds of methane and ethane. Perhaps the most widely used of these in industry is ammonia and in the household dichlorodifluoromethane ("Freon-12"). Nearly constant pressures are maintained on either side of the throttling valve by means of the compressor.

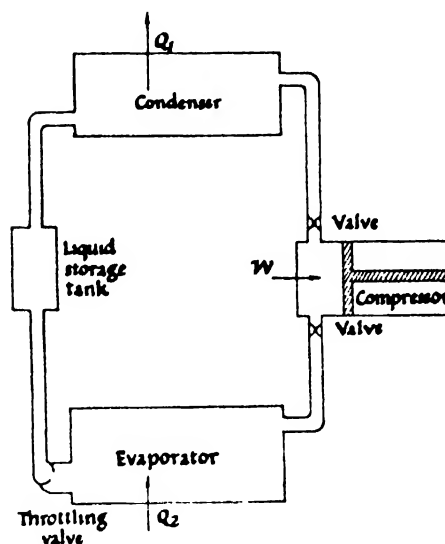


FIG. 1. Vapor-compression refrigerator from "Thermophysics," by Allen L. King, San Francisco, W. H. Freeman and Co., 1962).

The mixed liquid and vapor, which is colder than the near-surround, absorbs heat from the interior of the refrigerator box or cold room and completely vaporizes. The vapor then is compressed to a smaller volume by the compressor, and its temperature and pressure increase. Finally, the refrigerant pours through the outlet valve into the condenser, where it cools down and condenses to the liquid phase. Here heat is transferred either to cold air blown across the cooling coils or to cold water flowing by the cooling chamber of the condenser.

Comparative tests have shown that the coefficient of performance of vapor-compression refrigerators depends very little on the nature of the refrigerant. Because of mechanical inefficiencies, its actual value may be well below the ideal value—ordinarily, between 2 and 3. If a household refrigerator in which "Freon-12" is the refrigerant is operated between -15 and

30°C, its ideal coefficient of performance equals 4.8, but its actual value may be only 2.8.

Vapor-absorption Refrigerator. In this system there are no moving parts; the added energy comes from a gas or liquid fuel burner or from an electrical heater, as heat, rather than from a compressor, as work. A simplified diagram of it is shown in Fig. 2. The refrigerant is ammonia gas, which is liberated from a water solution and transported from one region to another by the aid of hydrogen. The total pressure throughout the system is constant and therefore no valves are needed.

Heat from the external source is supplied to the generator where a mixture of ammonia and water vapor with drops of ammoniated water is raised to the separator in the same manner as water is raised to the coffee in a percolator. Ammonia vapor escapes from the liquid in the separator and rises to the condenser, where it cools and liquefies. Before the liquefied ammonia enters the evaporator, hydrogen, rising from the absorber, mixes with it and aids in the evaporation process. Finally, the mixture of hydrogen and ammonia vapor enters the absorber, where water from the separator dissolves the ammonia. The ammonia water returns to the generator to complete the cycle. In this cycle heat enters the system not only at the generator but also at the evaporator, and heat leaves the system at both the condenser and the absorber to enter the atmosphere by means of radiating fins.

No external work is done, and the change in internal energy of the refrigerant during a complete cycle is zero. The total heat $Q_a + Q_c$ released to the atmosphere per unit time by the absorber and the condenser equals the total heat $Q_k + Q_e$ absorbed per unit time from the heater at the generator and from the cold box at the evaporator; so $Q_e = Q_a + Q_c - Q_k$, and therefore the coefficient of performance is $K = Q_e / Q_k = \{ (Q_a + Q_c) / Q_k \} - 1$.

The vapor-absorption refrigerator is free from intermittent noises; but it requires a continuous supply of heat, as from bottled gas or electrical generators. Refrigerators of this type are found in camps and farm houses not supplied with commercial electric power and in apartment houses where unnecessary noise is prohibited.

Thermoelectric Refrigerator. This device utilizes the thermoelectric effect first observed by Peltier in 1834 (see THERMOELECTRICITY). He discovered that heat is either absorbed or generated at the junction of two different conducting substances depending on the direction of an electrical current through it. In 1838, Lenz performed an experiment which may be considered the forerunner of thermoelectric refrigeration. He succeeded in freezing a drop of water at a bismuth-antimony junction by means of an electric current in one direction and then in melting it by reversing the current. However, not until 1909-11 were the properties of a thermocouple for refrigerating devices clearly specified. Then Altenkirch showed that its two conductors should have not only high thermoelectric coefficients but also high electrical

conductivities to reduce Joule heating and low thermal conductivities to minimize heat losses by conduction. Although semiconducting materials such as lead telluride or bismuth telluride have been found satisfactory, refrigerators incorporating them have only limited application because as yet they cannot compete economically with conventional mechanical refrigerators.

A modern thermoelectric refrigerator unit consists of one or more stages of series-connected *n*-type and *p*-type semiconductors as illustrated in Fig. 3(a). The charge carriers are electrons in the *n*-type semiconductor and holes in the *p*-type semiconductor. When a difference of potential is maintained across AB with B at the higher potential, the negative electrons carry both kinetic and potential energy away from C as they move toward B in the *n*-type semiconductor; the positive holes do the same as they move away from C toward A in the *p*-type semiconductor. Since energy is carried away from C to AB in both arms of the thermocouple, junction C becomes cold and junctions AB become warm. Temperature differences as high as 75°C have been obtained in a single-stage unit. Higher temperature differences may be produced by arranging several stages in cascade as illustrated by the two-stage system in Fig. 3(b). A small seven-stage thermoelectric refrigerator system only 1.5 inches high, employing *n*-type and *p*-type bismuth telluride alloys, was built at the research center of the Borg-Warner Corporation. With water at 27°C as the heat sink, its cold junction dropped to -118°C.

The coefficient of performance of a multistage thermoelectric refrigerator is no greater than that of one of its units. Let the temperature difference between AB and C in Fig. 3(a) be ΔT . As a result of the Peltier effect, the cold junction cools down at the rate $\bar{\alpha}(T_m - \frac{1}{2}\Delta T)I$, where $\bar{\alpha}$ is the mean value of the Seebeck coefficients for the *n*-type and *p*-type semiconductors, T_m is the mean temperature of the thermocouple, and I is the current through it. But due to Joule heating and thermal conduction from the warm to the cold junction, the rate of cooling at C is only $Q_c = \bar{\alpha}(T_m - \frac{1}{2}\Delta T)I - \frac{1}{2}I^2R - \lambda\Delta T$, where the total resistance of the couple $R = (l_p/A_p\sigma_p) + (l_n/A_n\sigma_n)$ and its thermal conductance $\lambda = (A_pk_p/l_p) + (A_nk_n/l_n)$ in which A_p and A_n are the cross-sectional areas of the *p*-type and *n*-type semiconductors, l_p and l_n are their lengths, σ_p and σ_n are their electrical conductivities, and k_p and k_n are their thermal conductivities. The power supplied externally must just equal the total Seebeck and Joule terms, namely, $P = \bar{\alpha}I\Delta T + I^2R$. The coefficient of performance, therefore, is

$$K = Q_c/P = \frac{\bar{\alpha}(T_m - \frac{1}{2}\Delta T)I - \frac{1}{2}I^2R - \lambda\Delta T}{\bar{\alpha}I\Delta T + I^2R}$$

It reaches a maximum value when the products of the thermal and electrical conductances for the two semiconductors have the same value and when the electrical resistance at the junctions is much smaller than R . The optimum current

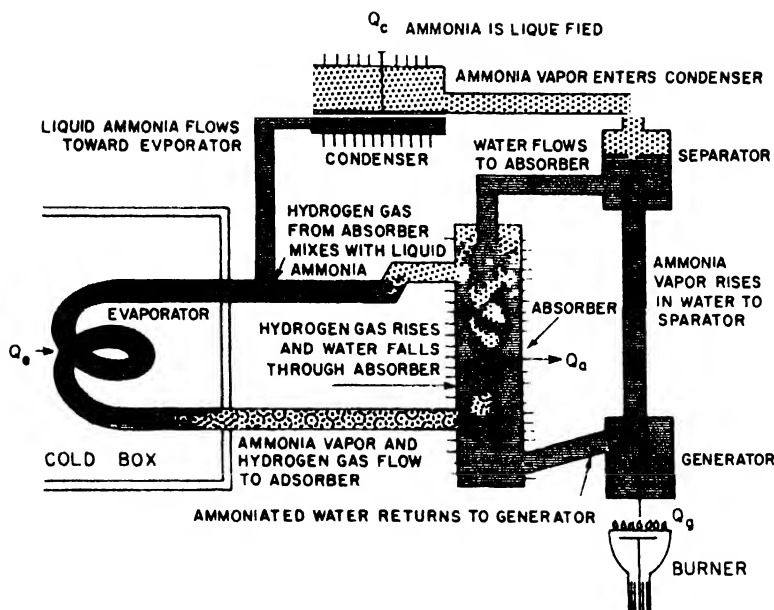


FIG. 2. Vapor-absorption refrigerator (from "Thermophysics," by Allen L. King, San Francisco, W. H. Freeman and Co., 1962).

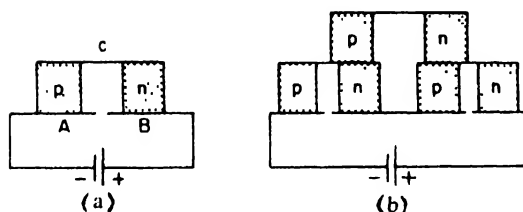


FIG. 3. Thermoelectric refrigerator: (a) single stage, (b) two-stage cascade.

then is given by the equation $(IR)_{opt} = \bar{\alpha} \Delta T [1 + ZT_m]^{-1}$ where Z is the *figure of merit* of the thermocouple,

$$Z = \bar{\alpha}^2 / [(k_p / \sigma_p)^{-1} + (k_n / \sigma_n)^{-1}]^2$$

The maximum value of K can now be written in the form

$$K_{max} = \frac{T_m (\sqrt{1 + ZT_m} - 1)}{\Delta T (\sqrt{1 + ZT_m} + 1)} - \frac{1}{2}$$

Thus K_{max} increases with an increase in the figure of merit, reaching the value $(T_m - \frac{1}{2} \Delta T) / \Delta T$ for very large Z . This is the coefficient of performance for an ideal thermodynamic machine.

The thermoelectric refrigerator is useful where space is at a premium or where, for other reasons, mechanical refrigerators are inconvenient, such as for spot-cooling electronic devices and for portable ice-cube makers.

Thermomagnetic and Magnetothermoelectric Refrigerators. In 1958 O'Brien and Wallace suggested that by means of the Ettinghausen effect, one should be able to achieve cooling for refrigeration purposes (see HALL EFFECT AND

RELATED PHENOMENA). This suggestion has been followed up with some success at low temperatures and even with the heat sink near room temperature by a special "cascading" device. In 1962, Smith and Wolfe discovered that the thermoelectric figure of merit for bismuth-antimony alloys can be increased by means of a magnetic field and that this enhancement is especially pronounced at low temperatures. These effects may be applied advantageously to refrigeration at low ambient temperatures.

ALLEN L. KING

References

- Goldsmid, H. J., "Applications of Thermoelectricity," London, Methuen and Co., Ltd., 1960.
- King, A. L., "Thermophysics," San Francisco, W. H. Freeman and Co., 1962.
- Wolfe, R., "Magnetothermoelectricity," *Sci. Am.*, 210, 70 (1964)
- Worthing, A. G., and Halliday, D., "Heat," New York, John Wiley & Sons, 1948.
- Zemansky, M. W., "Heat and Thermodynamics," New York, McGraw-Hill Book Co., 1957.

Cross-references: CRYOGENICS, HALL EFFECT AND RELATED PHENOMENA, HEAT, HEAT TRANSFER, LIQUEFACTION OF GASES, SEMICONDUCTORS.

RELATIVITY

The basic ideas of modern relativity theory are largely due to one man: Albert Einstein (1879-1955). Both main branches of pre-Einstein physics had relied on an absolute space. To Newton this had served as the agent responsible for a particle's resistance to acceleration; to Maxwell—in the guise of an "aether"—it was

the carrier of electromagnetic stresses and waves. Relativity may be defined briefly as the abolition of absolute space. Special relativity (1905) abolished it in its Maxwellian sense, and general relativity (1915) in its Newtonian sense as well.

Before looking at the theoretical background of relativity, we shall mention some of its more striking practical implications. According to special relativity, for example, a rod moving longitudinally at speed v through an inertial frame is shortened, relative to that frame, by a factor $\gamma = (1 - v^2/c^2)^{-1/2}$, where c is the speed of light. This factor increases with v ; when v is as large as $\frac{1}{2}c$, γ is only 1.01, but at higher speeds it grows rapidly and becomes infinite when $v = c$. The rate of a clock moving at speed v is decreased by the same factor γ ; this is one aspect of the revolutionary prediction that time is not absolute and that, for example, after journeying at high speed through space, one could, upon return, find the world aged very much more than oneself. In fact, time and space become merged in a four-dimensional continuum in which neither possesses more absoluteness than, e.g., the x -separation between points in a Cartesian plane, which depends on the choice of axes. According to special relativity, time- and space-separations between events similarly depend on the choice of motion of the observer. The mass of a body moving at speed v is also increased by the factor γ and thus becomes infinite at the speed of light. This illustrates another prediction of the theory—that no body or physical effect can travel faster than light. But the single most important result of special relativity, in Einstein's opinion, was the equivalence of mass m and energy E according to the formula $E = mc^2$. Although the original impact of special relativity was mainly theoretical and philosophical, technology since 1905 has made such vast strides (atomic power, particle accelerators, Mössbauer effect, etc.) that today special relativity is one of the most practical and, at the same time, best verified branches of all physics.

The same cannot yet be said for general relativity, whose importance is still largely theoretical. But its ideas are hardly less striking. General relativity is the modern theory of gravitation. Like special relativity, it pictures the world as a four-dimensional space-time continuum, but according to general relativity this is curved by the matter present in it. Particles and light rays are postulated to move along geodesics ("straightest possible" curves) in this four-space, and all reference to an absolute three-space as the standard of non-acceleration has disappeared. It is one of the marvels of this theory that, starting from such utterly different premises it nevertheless reproduces within experimental accuracy almost all the well-established results of Newton's (inverse square) gravitational theory. In the few cases where its predictions differ to a presently measurable extent from Newton's (as for the advance of the perihelia of the planets), general relativity has been borne out by observation. Furthermore, general relativity first led to

the construction of many interesting cosmological models, such as closed and finite universes; it also implies the possibility of gravitational waves and thus the need to quantize the gravitational field.

The theoretical basis of special relativity is Einstein's *special principle of relativity* which asserts that *all* the laws of physics are equally valid in *all* inertial frames of reference. This is an extension to the whole of physics of a relativity principle which the laws of mechanics have long been known to obey. Newton, as Galileo before him, illustrated this with the familiar example of a ship, "where all motions happen after the same manner whether the ship is at rest or is carried forward in a right line." The reason why Einstein's principle was revolutionary is that the known properties of light seem to contradict it at once. In our quasi-inertial terrestrial reference frame (which was assumed to coincide more or less with Maxwell's aether), light is propagated rectilinearly in all directions at constant speed. This fact is often called the *law of light propagation*. The validity of this law in all inertial frames would imply, for example, that a light signal emitted at the instantaneous coincidence of two observers O and O' who are moving uniformly relative to each other, each fixed in an inertial frame, spreads spherically with both observers considering themselves to remain permanently at the center of the sphere. Hence a light signal would always recede from an observer at the same speed, no matter how fast he chases it. The adoption of the special principle of relativity together with the law of light propagation thus seems to lead to absurdities. But, in fact, this is not so: it merely leads to the downfall of the classical ideas of space and time. It was part of Einstein's genius to recognize that these ideas were dispensable.

Two types of argument can be made in support of Einstein's principle. The first is experimental: all experiments devised to discover the frame of Maxwell's aether, such as the well-known MICHELSON-MORLEY EXPERIMENT (1887), failed to give positive results, though such results would have been well within range of observability. The second argument is theoretical, and rests on the unity of physics. For example, mechanics involves matter, which is electromagnetically constituted; electromagnetic apparatus involves mechanical parts; and so forth. If, then, physics cannot be separated into strictly exclusive branches, it would seem unlikely that the laws of different branches should have different transformation properties.

Consider now two observers O and O' like the ones mentioned earlier, and a light signal emitted at their coincidence. If each observer remains at the origin of a Cartesian reference system and sets his clock to read zero when the signal is emitted, the events on the light front must satisfy both the equations

$$\begin{aligned}x^2 + y^2 + z^2 - c^2 t^2 &= 0 \\x'^2 + y'^2 + z'^2 - c'^2 t'^2 &= 0\end{aligned}\quad (1)$$

where primes distinguish the space and time coordinates used by O' from those used by O . Now suppose the two observers arrange their corresponding y and z axes to be parallel, and their x axes to coincide. In classical mechanics, with this configuration of reference systems, the so-called *Galilean transformation equations*

$$x' = x - vt, \quad y' = y, \quad z' = z, \quad t' = t \quad (2)$$

relate the corresponding coordinates of any event. But under this transformation, the two equations (1) are not equivalent. Einstein showed that for these equations to be equivalent, the transformation equations must necessarily be

$$x' = \gamma(x - vt), \quad y' = y, \quad z' = z,$$

$$t' = \gamma(t - vx/c^2) \quad (3)$$

where $\gamma = (1 - v^2/c^2)^{-1/2}$. These are the well-known *Lorentz equations* which constitute the mathematical core of the special theory of relativity. They replace equations (2), to which they nevertheless approximate when v is small. The most striking of equations (3) is the last. It implies that events with the same value of t do not necessarily correspond to events with the same value of t' , which means that *simultaneity is relative*. Setting $x = 0$ in that equation also shows that the clock at the origin of O goes slow by a factor γ in the frame of O' . But, setting $x = vt$, we see that the clock at the origin of O' similarly goes slow in the frame of O . Setting $t = 0$ in the first of equations (3), we see that a rod, fixed in the frame of O' along the x' axis, appears shortened by a factor γ in the frame of O ; this phenomenon too can be shown to be symmetric between the frames.

Another important property of equations (3) is that they leave invariant the differential quadratic

$$ds^2 = dx^2 + dy^2 + dz^2 - c^2 dt^2 \quad (4)$$

which leads to the possibility of mapping events in a four-dimensional pseudo-Euclidean *space-time* in which an absolute *interval* ds exists, and in which the language and results of four-dimensional geometry can thus be applied. For example, a uniformly moving particle is described simply by a straight line in this space-time.

Since Newton's laws of mechanics are invariant under the transformation (2) and *not* (3), it was necessary to amend these laws so as to make them "Lorentz invariant." It was found possible to do this by retaining the classical laws of conservation of mass and momentum but postulating that the mass of moving bodies increases by the factor γ , a fact amply borne out by modern particle accelerators. This led to the theoretical discovery of the equivalence of mass and energy—most spectacularly exemplified by the atomic bomb.

In contrast to Newton's theory, Maxwell's vacuum electrodynamics was compatible with Einstein's theory. Lorentz, independently of Einstein, but without realizing the full significance of his result, had already discovered equations (3)

as precisely those which leave Maxwell's equations invariant. In other words, Maxwell's equations already were "Lorentz invariant" and needed no modification. Nevertheless relativity has considerably deepened our understanding of Maxwell's theory. Other branches of physics, like kinematics, optics, hydrodynamics, thermodynamics, non-vacuum electrodynamics, etc., all underwent slight modifications to make them Lorentz invariant. Only Newton's inverse square gravitational theory proved refractory; several Lorentz invariant modifications of it were proposed but none were entirely acceptable.

Einstein eventually solved the gravitational problem in an unexpected way. He rejected Newton's absolute space as the cause of inertia on the grounds that "it is contrary to the spirit of science to conceive of a thing which acts but cannot be acted upon." His general theory of relativity ascribes to the space-time continuum discovered by special relativity the role of an inertial guiding field (free particles and light follow geodesics) but allows this field to be affected (curved) by the matter in it.

This extension was made possible by the so-called *principle of equivalence*. To Newton, an inertial frame was, primarily, the frame of "absolute space" in which the stars were assumed to be fixed, and, secondarily, any frame moving uniformly relative to absolute space. Thus an inertial frame exhibited its defining property, viz., that in it free particles move uniformly and rectilinearly (Newton's first law), only in the regions far from attracting masses. In 1907 Einstein changed this global definition to a local one: a local inertial frame is a freely falling non-rotating reference system. (The meaning of "local" is here determined by the extent to which the nonuniformity of the gravitational field is negligible.) Within the limits of each such frame Newton's laws of mechanics would be valid according to the classical theory; in particular, Newton's first law would be strictly satisfied. Now Einstein once again made the generalization from mechanics to the whole of physics. His principle of equivalence asserted that all the laws of physics are the same in each local inertial frame. It is these frames, therefore, which are the proper province of the special principle of relativity. Special relativity now becomes a local theory. In recompense, we need no longer go to the tenuous interstellar regions for its strict validity.

As we have seen, special relativity forces a four-dimensional metric structure [Eq. (4)] on the events within an inertial frame. By patching together the structures of all the local inertial frames, we obtain the structure of the world of general relativity. Locally, it can be regarded as flat. But it is evident that, if the very suggestive geodesic law of motion is to hold, the presence of matter must impress a curvature on this space-time. For example, the planets move in patently curved paths around the sun (in four-dimensions these are helicoidal rather than elliptical); for these paths to be geodesic, the space-time around

the sun must be curved. Just how matter curves the surrounding space-time is expressed by Einstein's field equations

$$G_{ij} = -\frac{8\pi G}{c^4} T_{ij} \quad (5)$$

which look deceptively simple. Technically, they represent ten second-order partial differential equations for the metric of space-time. This metric enters the 16 components of the "Einstein tensor" G_{ij} , of which only 10 are independent, for $G_{ij} = G_{ji}$. G is the constant of gravitation; T_{ij} is the so-called energy tensor of the matter, and its components represent a generalization of the classical concept of density.

The exact solution of Eq. (5) has been possible only in a limited number of physical situations. For example, in 1916 Schwarzschild gave the exact solution for the space-time around a spherical mass m (e.g., the sun):

$$ds^2 = (1 - a/r)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) - (1 - a/r)c^2 dt^2 \quad (6)$$

where $a = 2Gm/c^2$, r is a measure of distance from the central mass, t is a measure of time, and θ and ϕ are the usual angular coordinates. Note that when $m = 0$, Eq. (6) reduces simply to the flat space-time of Eq. (4), written in polar coordinates, and its geodesics would be straight lines (in space and time). But for Eq. (6), the geodesics in the plane $\theta = \pi/2$ are found to satisfy the equation

$$\frac{d^2u}{d\phi^2} + u = \frac{Gm}{h^2} + \frac{3Gmu^2}{c^2} \quad (7)$$

where $u = 1/r$ and h is a constant. This differs formally from the classical orbit equation only by the presence of the last term, which is very small. But as a consequence of that term, the solution of Eq. (7) is

$$u = Gmh^{-2}(1 + e \cos p\phi), \quad p = 1 + 3G^2m^2h^{-2}c^{-2} \quad (8)$$

instead of the classical solution which has $p = 1$. Now r is a function in ϕ of period $2\pi/p$ instead of 2π , and therefore the orbital ellipse precesses. For the planet Mercury, for example, the secular precession predicted is $42''$ (seconds of arc), and this agrees well with observation. In the space-time defined by Equation (6) one also finds that light-signals which pass close to the central mass are bent by an angle twice as big as that predicted on a simple Newtonian corpuscular theory of light; and again observations bear out the relativistic prediction. The third "crucial" prediction, which has also been verified observationally, is the reddening of the light received from the surface of very dense stars.

W. RINDLER

References

- Bergmann, P. G., "Introduction to the Theory of Relativity," New York, Prentice-Hall, 1942.
 Eddington, A. S., "Space, Time, and Gravitation, Cambridge, The University Press, 1920.
 Einstein, A., *et al.*, "The Principle of Relativity," New York, Dover, 1923.
 Møller, C., "The Theory of Relativity," Oxford, Clarendon Press, 1952.
 Pauli, W., "Theory of Relativity," London, Pergamon Press, 1958.
 Rindler, W., "Special Relativity," New York, John Wiley & Sons, 1960.
 Synge, J. L., "Relativity: The Special Theory," Amsterdam, North Holland, 1956; "Relativity: The General Theory," Amsterdam, North Holland, 1960.
 Tolman, R. C., "Relativity, Thermodynamics, and Cosmology," Oxford, Clarendon Press, 1934.

Cross-references: ACCELERATORS, PARTICLE; ATOMIC ENERGY; DYNAMICS; MICHELSON-MORLEY EXPERIMENT; MOSSBAUER EFFECT; TIME; VELOCITY OF LIGHT.

RELAXATION

By relaxation is understood the phenomenon that an observable time elapses between the moment when a system in equilibrium is subjected to a momentary change in condition and the moment when the system is again in equilibrium.

A good example of a system showing relaxation is an uncharged capacitor with a capacitance C connected in series with a resistance R , to which circuit a constant voltage U is applied at a time $t = 0$. The charge Q of the capacitor at time t is given by the differential equation

$$U = \frac{Q}{C} + R \frac{dQ}{dt} \quad (1)$$

the solution of which is:

$$Q = CU(1 - e^{-\frac{t}{RC}}) \quad (2)$$

Hence the charge Q does not follow the sudden change of U but shows an exponential increase towards the final value CU (see Fig. 1). The time $\tau = RC$ is referred to as the relaxation time of the system. Obviously τ governs both the rate of charging and the rate of discharge: if the capacitor with a charge Q_0 is shortcircuited via a resistance at the time $t = 0$, then the charge at time t is given by:

$$Q = Q_0 e^{-\frac{t}{\tau}}$$

Relaxation is met with in various fields of physics, e.g.:

(a) If a spring connected in parallel with a dashpot is suddenly loaded with a constant force, it will take some time until this system is again in equilibrium.

(b) If an electrically or magnetically polarizable substance is suddenly placed in a constant

electric or magnetic field, it will take some time until the electric or magnetic dipoles have orientated themselves in the field.

(c) If the irradiation of a phosphorescent substance is suddenly stopped, the phosphorescence does not cease immediately but continues for some time during which the intensity decreases exponentially.

An electrical analogy can be devised for all these cases. For instance, if in case (a) we substitute a voltage for the force, a charge for the displacement, the reciprocal of a capacitance for the spring constant f , and a resistance for the coefficient of friction η , we obtain exactly the differential equation [Eq. (1)]. The relaxation time τ will then be $\tau = \eta/f$, which quantity depends on temperature since as a rule η greatly varies with temperature. Generally, the relaxation times in electrically and magnetically polarizable media also vary with temperature.

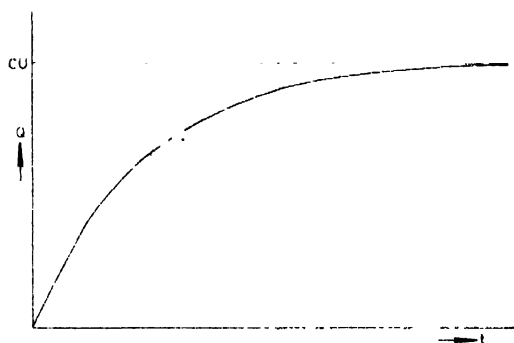


FIG. 1. Q vs t .

If we subject a system having relaxation properties to a periodically varying change in condition instead of one sudden change in condition, some characteristics are revealed which are likewise typical of relaxation. Let us again take the RC circuit as an example, this time applying to it a voltage $U = U_0 \sin \omega t$. Solution of the differential equation [Eq. (1)] yields the following expression for the charge Q of the capacitor:

$$Q = \frac{CU_0}{\sqrt{1 + \omega^2\tau^2}} \sin(\omega t - \text{tg}^{-1} \omega\tau)$$

The charge has a component Q' which is in phase with the voltage applied and a component Q'' which is shifted 90° in relation to the voltage applied. Q' and Q'' , usually referred to as the real and the imaginary component of Q , respectively, are given by:

$$Q' = CU_0 \frac{1}{1 + \omega^2\tau^2}$$

$$Q'' = CU_0 \frac{\omega\tau}{1 + \omega^2\tau^2}$$

The variation of Q' and Q'' with $\omega\tau$ is shown in Fig. 2 where $\omega\tau$ has been plotted logarithmically in order to obtain symmetrical curves.

Q'' is a measure of the energy dissipated in the circuit per cycle; it has its maximum value at $\omega = 1/\tau$ and a negligibly small value at much higher and much lower frequencies.

Furthermore, Q'' reaches half its maximum value at $\omega = 3.73/\tau$ and $\omega = 0.27/\tau$ so that the half-width value $\Delta\omega$ of the Q'' curve expressed in angular frequency units is $3.46/\tau$. At $\omega = 1/\tau$, Q' just has half its maximum value.

Hence, by measuring Q' and Q'' vs frequency, we have various possibilities for determining the relaxation time.

If a dielectric solid is subjected to a periodically changing electric stress and the real (in-phase) and imaginary (out-of-phase) components of the dielectric constant of the substance are determined, curves will be found having a shape as shown in Figs. 2(a) and 2(b), from which the relaxation time can be determined. The mechanical relaxation time of an elastic solid can be determined in a similar manner.

In some cases, not one but several absorption peaks with the corresponding changes of the real component are found in this kind of experiment, from which it can be concluded that several relaxation mechanisms are involved. This phenomenon is frequently observed in polymers where the various components, such as short chains, long chains, cross-linked chains, etc., have

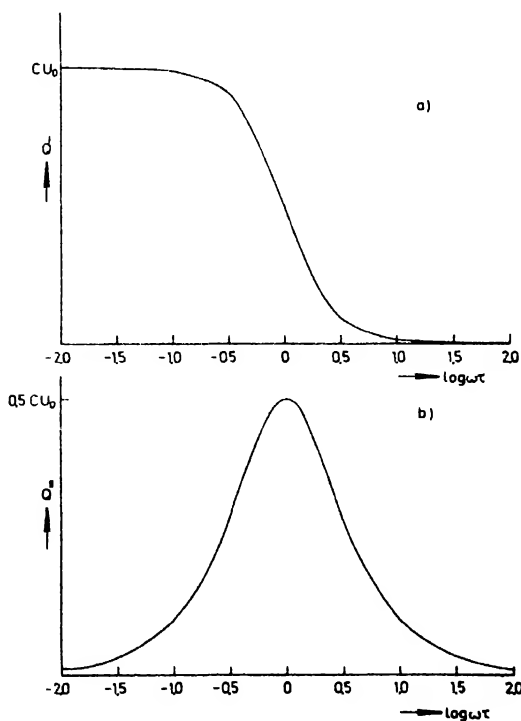


FIG. 2. Q' (a) and Q'' (b) vs $\log \omega\tau$.

different possibilities of moving and therefore have different relaxation times. There are also cases in which there is a more or less continuous distribution of relaxation times, and consequently, an absorption which is practically independent of frequency.

Since the above-mentioned electric circuit has no inductance (which corresponds to the masses of the particles in magnetic, electric and mechanical systems), there can be no resonance either.

As an example of a system having both resonance and relaxation properties, we will consider a substance containing magnetic nuclei, e.g., hydrogen nuclei, which at the time $t = 0$ is placed in a magnetic field H . In this substance a magnetic moment M is built up in the direction of H by the fact that the original random orientation of the magnetic dipoles of the nuclei changes under the influence of the magnetic field into a distribution of such a nature that a resulting moment M arises in the direction of the magnetic field. M is determined by an equation analogous to Eq. (2), namely:

$$M = M_0(1 - e^{-\frac{t}{T_1}})$$

where M_0 is the final value of M and T_1 is a relaxation time known as longitudinal or spin-lattice relaxation time, which is associated with the interaction between the magnetic dipoles and their surroundings.

Now, if the direction of the magnetic field is suddenly changed, M will make a precessional motion about the new H due to the fact that the resulting angular momentum J is coupled to the magnetic moment M of the nuclei via the relation $\gamma = M/J$ where γ is the magnetogyric ratio of the nuclei.

The angular frequency of the precessional motion is given by:

$$\omega = \gamma H$$

On behalf of the precessional motion, the magnetic moment M has not only a longitudinal component M_1 in the direction of the new magnetic field but also a transverse component M_2 at right angles to the new magnetic field. Whereas it would seem logical to expect that M_1 would increase and M_2 decrease exponentially with the relaxation time T_1 due to the relaxation mechanism, in reality this is not so.

Experiments have shown that M_2 decreases faster than the increase of M_1 so that apart from the longitudinal relaxation time T_1 there must also be a separate transverse relaxation time T_2 which invariably is smaller than or equal to T_1 . The existence of a T_2 besides T_1 can be explained by the fact that apart from the return of individual magnetic moments of the nuclei to the direction of H (which is characterized by T_1), it is also possible for these individual magnetic moments—which together form M —to continue their precessional motion along the same conical shell though getting out of phase by mutual interaction. In the latter case, M_2 decreases whereas M_1 remains unchanged.

Just as with nonresonating systems, periodically changing quantities can be successfully introduced in resonating systems. In the case of a substance containing magnetic nuclei, this is done by applying, beside a stationary magnetic field H_z in the z -direction, also a periodically changing magnetic field $2H_1 \sin \omega t$ in the x -direction. This is known as a nuclear magnetic resonance experiment. In this case, a periodically changing magnetization intensity is created in the x -direction whose real (in-phase) component M' and imaginary (out-of-phase) component M'' are given by:

$$M' = \gamma M_0 H_1 \frac{T_2^2(\gamma H_z - \omega)}{1 + T_2^2(\gamma H_z - \omega)^2 + \gamma^2 H_1^2 T_1 T_2}$$

$$M'' = \gamma M_0 H_1 \frac{T_2}{1 + T_2^2(\gamma H_z - \omega)^2 + \gamma^2 H_1^2 T_1 T_2}$$

where M_0 is the magnetic moment when no periodically changing magnetic field is applied.

These quantities can both be determined experimentally. Usually the experiment is done so that ω is maintained constant and H_z is slowly varied. In this case M'' is proportional to the energy absorbed by the system per unit of time.

The variation of $M'/\gamma M_0 H_1 T_2$ and $M''/\gamma M_0 H_1 T_2$ has been plotted against $T_2(\gamma H_z - \omega)$ in Fig. 3, it having been assumed that H_1 is so small that $\gamma^2 H_1^2 T_1 T_2$ is much smaller than 1. It will be seen that the curves differ considerably from those in

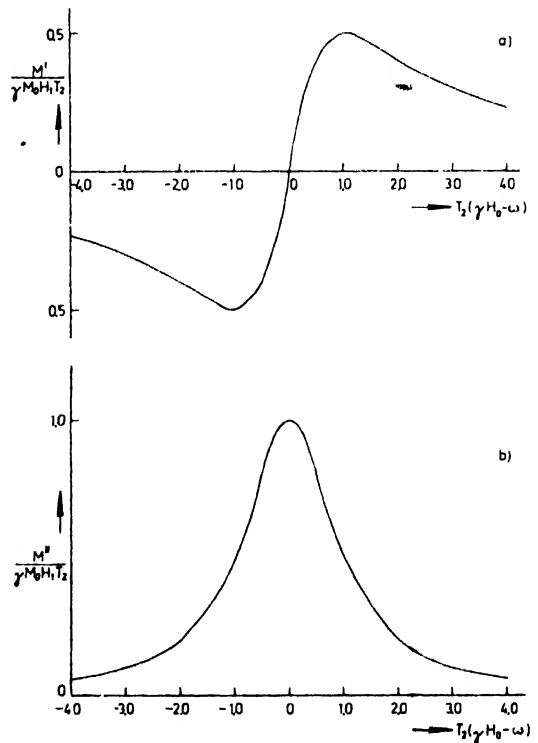


FIG. 3. Shape of the dispersion (a) and absorption (b) curves.

Fig. 2. The absorption curve has its maximum at the value $\gamma H_z = \omega$ which value has nothing to do with the relaxation times but only with the concurrence of the precessional frequency γH_z of the magnetic moment in the magnetic field H_z with the frequency of the periodically changing magnetic field. Hence it is clearly a matter of resonance absorption. The half-width value $\Delta\omega$ of the absorption curve, expressed in terms of a frequency, is $2/T_2$. The $M'/\gamma M_0 H_1 T_2$ curve is the curve of anomalous dispersion, which at $\gamma H_z = \omega$ has just the value 0, which is invariably found with resonating systems. With increasing values of H_1 both curves get wider and wider and the absorption curve eventually becomes 0 throughout. This phenomenon is called saturation. T_1 and T_2 can be calculated in principle from curves plotted for different known values of H_1 . Both relaxation times are very much dependent on temperature and on the phase of the substance under test.

The same phenomena are observed in the case of resonance of unpaired electrons, known as electron spin resonance.

Generally speaking, the investigation of relaxation times provides information on the surroundings of relaxing particles and thus can contribute to our knowledge about the structure of substances.

Furthermore, profound knowledge of relaxation times is essential in many instances to the control of physical processes (e.g., the creation of low temperatures by magnetic means) and physical techniques (e.g., the maser technique).

J. SMIDT

References

- Abragam, A., "The Principles of Nuclear Magnetism," London, Clarendon Press, 1961.
 Bottcher, C. J. F., "Theory of electric polarization," New York, Elsevier, 1952.
 von Hippel, A. R., "Molecular Science and Molecular Engineering," Cambridge, Mass., University Technical Press M.I.T., 1959.
 Laukien, G., in Flugge, S., Ed., "Handbuch der Physik," Part 38/1, p. 131, Berlin, Springer, 1961.
 Mason, P., in Flugge, S., Ed., "Handbuch der Physik," Part 11, p. 361, Berlin, Springer, 1961.

Cross-references: ABSORPTION SPECTRA; CAPACITANCE; CONDUCTIVITY, ELECTRICAL; ELECTRON SPIN; LUMINESCENCE; MAGNETIC RESONANCE; MAGNETISM; RESONANCE.

REPRODUCTION OF SOUND

History. In 1807 the British physicist Thomas Young designed the first device capable of making a graphic record of sound waves. His description of the principle of sound recording is as clear and valid today as it was then, and it may serve here:

"The situation of a particle at any time may be represented by supposing it to mark its path, on a surface sliding uniformly along in a transverse direction. Thus, if we fix a small pencil in a vibrating rod, and

draw a sheet of paper along, against the point of the pencil, an undulated line will be marked on the paper, and will correctly represent the progress of the vibration."

The recording stylus of Young's device had to be touched directly with the sound source. The "phonautograph" developed by Léon Scott de Martinville in 1856 was able to record sound from the air, via a horn, parchment diaphragm, and hog-bristle stylus. Both the Young and Scott devices made a helical trace on a rotating cylinder. Neither recording could be played back because the recorded trace was not deep or stiff enough to guide the vibrations of a reproducing stylus.

In April 1877, Charles Cros deposited with the French Academy of Sciences a sealed package containing the description of a complete record-reproduce system. Cros planned to use metal records photoengraved from an original tracing in lampblack, but never carried out his plans. In the fall of that year, Edison constructed a "phonograph" whose recording stylus made indentations on tinfoil wrapped on a pre-grooved cylinder. Although these indentations were partly deformed by the playback stylus, a weak, distorted, but intelligible version of the human voice could be reproduced.

Modern disc recording is closer to the Cros system than to Edison's; commercial recordings are stampings from a hardened mold. Virtually all original recordings for home use are first made on magnetic tape. They are then cut into a master disc made of a relatively soft material, lacquer, and end up as a metal stamper from which mass impressions can be made in a vinyl composition.

Modern Recording Systems. Current recording media include magnetic tape, transparent film, and grooved discs. The most widely used is the disc.

If the disc cutter head is fed with constant electrical energy over the frequency spectrum, it will produce constant average velocity in the recording stylus. Since the wave length of the recorded signal is doubled with each lower octave, constant stylus velocity would produce impractically large groove excursions in the bass range. Progressive bass attenuation is therefore introduced below a frequency which has now been standardized at 500 cycles, called the *turnover* frequency, in such an amount that the amplitude of groove modulation for signals below this frequency remains the same at a given power. A compensatory boosting of the bass frequencies must be employed in the playback amplifier.

A second problem in disc recording has to do with surface noise introduced by irregularities in the record material and picked up by the needle. The noise is distributed fairly evenly over the frequency spectrum on the basis of energy per cycle. Since each higher octave covers twice the number of cycles, this noise may be considered primarily a treble phenomenon (see NOISE, ACOUSTICAL).

The signal to the recording cutter is again

altered, this time by progressive treble boost. In playback, a compensatory treble attenuation is introduced which brings the recorded signal back to normal and at the same time significantly reduces the amount of surface noise. This system of recording *preemphasis* does not change the treble content of the final reproduced sound, but it increases the amplitude of the high-frequency groove modulations relative to the random surface irregularities in the recorded material.

These changes of frequency balance in the recorded signal are called *equalization*; the particular equalization curve is called the *recording characteristic*.

The Modern Disc Reproducing System. A pickup, also called a cartridge, traces the groove modulations through a needle or stylus, whose vibrations are converted to an electrical signal. The most common types of electrical generator employed in cartridges are the ceramic (piezo-electric), moving magnet, moving coil, and variable reluctance. It is the task of the cartridge to translate faithfully the wave forms of the groove into an electrical signal. Some of the problems in cartridge design have to do with the stiffness of the moving needle system, which tends to wear the record, and with the mass of the needle tip, which tends to resonate with the semi-elastic record material at high frequencies.

The turntable must revolve the record at a constant speed. Periodic variations in this speed are called *flutter* (the onomatopoeic term for very slow flutter is *wow*). Any noise introduced into the signal by the moving parts (via the pickup) is called *rumble*. A pivoted arm holds the cartridge in place over the record groove, with the vibration axis of the cartridge approximately tangent to the groove over the entire radius of the record.

The amplifier, which may be on one or more chassis, has two sections: the *preamplifier* and the *power amplifier*. The former is a voltage amplifier which performs the functions of input program selection, tone control, volume control, compensation for the frequency equalization introduced in recording, and voltage amplification of the input signal to the point where it can drive the power amplifier. The latter builds up the electrical signal power so that the signal is able to drive the loudspeaker or speakers.

An amplifier is a device whose output energy is greater than, but in the same form as, the input signal energy. The amplifier must therefore borrow energy from an outside source, normally the electrical power line. There are many types of basic amplifying devices. The two that are used almost universally in sound reproduction are the vacuum tube and the transistor.

The loudspeaker converts the electrical output of the amplifier into acoustical energy, usually through a vibrating diaphragm. The loudspeaker and the cartridge, because they are mechanical devices with their own resonances and characteristic behavior, are more intransigent to precise control by the input signal than electronic circuits.

A reproducing system is designed for minimum noise, distortion, frequency discrimination over the audible spectrum, and transient ringing; speakers must also have adequate treble dispersion.

Stereophony. Stereophonic sound is recorded on two separate channels from separate microphone inputs. Just as each lens of a stereoscopic camera takes a complete picture, each microphone channel picks up all of the sound, one from a right-oriented perspective and the other from a left-oriented perspective. When the two channels are played back through separate right and left speakers, the sense of the acoustical atmosphere of the concert hall is enhanced; a corollary of this is an increased clarity of inner melodic voices. The ability of the listener to determine the apparent position of different musical instruments is a less important part of the stereo effect.

In stereo disc recording, one channel is, in effect, recorded on the left groove wall, and the other channel on the right groove wall. The reproducing stylus must execute a complex motion containing both vertical and horizontal components in order to follow the modulations of both groove walls simultaneously. These complex movements are analyzed into two vectors by two separate generating elements, each at 45° to the vertical and on opposite sides. Each generating element produces the electrical signal for one channel.

The ultimate design goal of sound reproducing equipment is not to create "better" sound, but to efface the imprint of the equipment, so that the original musical quality is recreated.

EDGAR R. VILLCHUR

References

- Beranek, Leo L., "Acoustics," New York, McGraw-Hill Book Company, Inc., 1954.
- Hunt, Frederick, V., "Electroacoustics," Cambridge, Harvard University Press, 1954.
- Olson, Harry F., "Acoustical Engineering," New York, D. Van Nostrand Company, 1957.
- Villchur, Edgar, "Reproduction of Sound," Cambridge, Acoustic Research, Inc., 1962.
- Villchur, Edgar, "Reproduction of Sound," *Phys. Today*, 5, No. 9 (September, 1952).

Cross-references: ACOUSTICS; ARCHITECTURAL ACOUSTICS; MUSICAL SOUND; NOISE, ACOUSTICAL; PHYSICAL ACOUSTICS.

RESISTANCE. See CONDUCTIVITY, ELECTRICAL.

RESONANCE

The phenomenon which scientists call resonance can be identified in many different physical systems of widely varying sizes. For instance, a father who pushes his small child on a swing finds that with each successive push the swing

goes higher. An astronomer who examines the spectrum of the sun notes the appearance of a series of dark lines superposed on the continuous red to violet band of colors. The solid state physicist who observes the amount of electromagnetic radiation transmitted through a waveguide at a particular microwave frequency finds it can be sharply reduced if certain materials are placed in the guide and a magnetic field is applied. All of these effects are examples of resonance, yet the actual physical mechanisms are different and the explanations require different analytical procedures depending on whether the system is governed by classical or quantum theory.

In the case of the father pushing the swing, the resonance can be explained as a direct consequence of Newton's laws of classical mechanics. In simple terms, it occurs when an outside agent (the father) pushes the system (swing with child) with a periodic force having the same frequency as the natural frequency of the system itself. The natural frequency is the frequency with which the system oscillates if it is displaced from its normal position of equilibrium and then released to swing freely back and forth. If the external push is timed so as to be exactly in step with the natural frequency, one can think of the swing as being steadily accelerated in its natural direction of motion. When the natural direction of motion changes at the end points of the path so does the push. Elementary kinematics then predicts that such a steady acceleration will cause both the displacement and the maximum velocity to increase continuously and eventually become very large. It is this extreme magnitude of displacement and velocity which is the most noticeable aspect of resonance in a mechanical system. However, less apparent visually, yet equally important, is the rate of energy transfer between the source and the system at resonance. The external agent transfers energy in the form of mechanical work to the system. At resonance, the average rate of this energy transfer per cycle becomes a maximum. This property of maximum rate of energy transfer is of great value when analyzing the effect of radiation on microscopic systems such as atoms and nuclei whose physical behavior must be explained by the laws of quantum physics rather than classical physics. In these systems as in mechanical systems a necessary condition for resonance is that the frequency of the radiation must match a frequency which is in some way associated intrinsically with the system. But here the coordinates of position and velocity are no longer suitable ones to use in describing the response at resonance. Nevertheless, the resonant system is still characterized by the general property that the rate of energy transfer into it is a maximum. In order to bring out the essential features of resonance involving both the behavior of position and velocity and the rate of energy transfer, it is simplest to examine a mechanical system consisting of a mass on a spring with frictional damping.

Resonance in a Simple Mechanical Oscillator. We begin with a mass M attached to a fixed point

by a spring with a force constant K and a damping resistance R . In addition the mass is acted upon by a periodic external force of frequency ν , $F = F_0 \cos 2\pi\nu t$. The displacement of the mass from its normal equilibrium position is represented by x . The acceleration d^2x/dt^2 of the mass M is determined by the resultant of these three forces, all along the x -direction: $-Kx$, the elastic restoring force; $-R(dx/dt)$, the damping force proportional to the instantaneous velocity dx/dt ; and the external force $F_0 \cos 2\pi\nu t$. Newton's second law of motion, which states that resultant force = mass \times acceleration, can be expressed by a second-order inhomogeneous linear differential equation:

$$M \frac{d^2x}{dt^2} + R \frac{dx}{dt} + Kx = F_0 \cos 2\pi\nu t \quad (1)$$

However, in order to discuss resonance it is first necessary to describe the behavior of the spring system when no external driving force is being applied. If the mass is displaced from its normal equilibrium position and then released, it will move in simple harmonic motion with the natural frequency ν_1 given by

$$\nu_1 = \frac{1}{2\pi} \sqrt{\frac{K}{M} - \frac{R^2}{4M^2}}, \quad \left(\frac{K}{M} - \frac{R^2}{4M^2} \right) \quad (2)$$

This result can be derived mathematically by setting the term on the right-hand side of Eq. (1) equal to zero thus making the equation homogeneous. It can be solved for $x(t)$ by standard methods (see references 1 and 2). The solution is:

$$x = Ae^{-\frac{R}{2M}t} \sin(2\pi\nu_1 t + \phi) \quad (3)$$

where A and ϕ are constants depending on the initial position and velocity. For example, if the mass starts from a position $x = x_0$, with zero velocity, then $A = x_0$ and $\phi = 90^\circ$, and Eq. (3) becomes

$$x = x_0 e^{-\frac{R}{2M}t} \cos 2\pi\nu_1 t \quad (4)$$

Since the solution is the product of a negative exponential function and a sinusoidal function, the amplitude of the displacement will diminish a little bit more with each succeeding cycle of oscillation and eventually the displacement becomes zero again. This is called a damped harmonic oscillation and is physically what we see any time a real spring or pendulum is disturbed and then allowed to oscillate freely. The damping of the oscillations can also be looked upon as representing a conversion of mechanical energy into heat which is proceeding at the instantaneous rate $R(dx/dt)^2$.

We now consider the solution of the inhomogeneous equation [Eq. (1)]. From the theory of linear differential equations, the solution of an inhomogeneous equation can be expressed as the sum of solutions of the homogeneous and inhomogeneous equation. Thus a general solution can be written down in which one term is the

solution [Eq. (3)] while the second is the particular integral satisfying Eq. (1).

$$x = Ae^{-\frac{R}{2M}t} \sin(2\pi\nu_1 t + \phi) + \frac{F_0}{2\pi M} \frac{\cos(2\pi\nu t + \alpha)}{\sqrt{4\pi^2(\nu_0^2 - \nu^2)^2 + \frac{\nu^2 R^2}{M^2}}} \quad (5)$$

where $\nu_0 = (1/2\pi) \sqrt{K/M}$ is the natural frequency for the undamped oscillator and α , the phase angle between the applied force and the displacement, is defined by

$$\tan \alpha = \frac{\nu R}{2\pi M(\nu_0^2 - \nu^2)} \quad (6)$$

Since the first term in Eq. (5) becomes vanishingly small and can be neglected after a period of time has elapsed, it is called the transient part of the solution. The particular integral maintains a constant amplitude as long as the driving frequency does not change, and it is called the steady-state part of the solution.

The steady-state solution for the instantaneous velocity is also of great interest

$$v = \frac{dx}{dt} = \frac{F_0\nu}{M} \frac{\cos(2\pi\nu t + \beta)}{\sqrt{4\pi^2(\nu^2 - \nu_0^2)^2 + \frac{R^2\nu^2}{M^2}}} \quad (7)$$

where β , the phase angle between applied force and velocity, is defined by

$$\tan \beta = \frac{2\pi M(\nu^2 - \nu_0^2)}{\nu R} \quad (8)$$

It is seen that both displacement and velocity have amplitudes which depend on the frequency of the applied force. The form of Eq. (5) and (7) is such that there must exist frequencies for which each achieves a maximum value. The frequencies which produce these maximum values can be found by differentiating the displacement or velocity function with respect to frequency, setting the derivative equal to zero, and solving for ν . In this way, it is deduced that the amplitude of x becomes a maximum when

$$\nu = \sqrt{\nu_0^2 - \frac{R^2}{8\pi^2 M^2}} \quad (9)$$

while the amplitude of v becomes a maximum when

$$\nu = \nu_0 \quad (10)$$

This latter frequency ν_0 which is the natural frequency of the undamped oscillator is customarily referred to as the resonance frequency of the oscillator. The phase difference between the force and velocity at resonance is zero. Thus the velocity will achieve its largest magnitude in the part of the cycle when the push is greatest, confirming what we sense intuitively in pushing a swing. The displacement resonance occurs at a

frequency slightly different from the resonance frequency, but the difference becomes smaller as R decreases. The velocity resonance is of significance in discussing energy because the rates of energy transfer into and out of the oscillator depend on velocity. The external periodic force is doing work on the oscillator at the instantaneous rate of Fv , while the oscillator is working against the damping force at the instantaneous rate of $-Rv^2$. Thus energy is being simultaneously absorbed and dissipated. Since the instantaneous rates vary periodically with time, it is more meaningful to calculate average rates per period and then compare. Such a calculation (see reference 2) shows that the average rate of energy absorption is exactly equal to the average rate of energy dissipation and has the magnitude $(F_0^2/2R) \cos^2 \beta$. At velocity resonance $\cos^2 \beta = 1$, and these rates of absorption and dissipation will have maximum magnitudes of $F_0^2/2R$.

Thus we find that a mechanical oscillator will resonate when a periodic external force with a frequency equal to the natural frequency of the oscillator acts on it. In addition, the rate of absorption of energy from the external source is a maximum at this same resonance frequency. This second property of maximum rate of energy absorption becomes useful in the discussion of resonance phenomena in systems where resonance cannot be described in terms of what happens to the state variables of position and velocity.

There is an exact analogue to this analysis in the electrical circuit consisting of an inductance L , a capacitance C , and a resistance R in series and driven by an alternating emf, $E_0 \cos 2\pi\nu t$. The differential equation for the variation of charge q with time is

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = E_0 \cos 2\pi\nu t \quad (11)$$

and resonance will occur under the same corresponding conditions as for the mechanical oscillator (see ALTERNATING CURRENTS).

Resonance Phenomena in Atomic-sized Systems.

The exact theory explaining the processes whereby systems composed of atoms absorb and emit energy is based on the principles of quantum theory. Nevertheless, such processes still show some physical analogy to resonance in a mechanical oscillator. A simple example is the single atom consisting of one or more electrons bound to a nucleus by predominantly electrostatic forces. Quantum theory predicts that the atom can exist only in a set of discrete energies in contrast to the continuous range of energies which in principle is available to a large-size satellite system such as the earth and moon. If the atom changes from a higher-energy state E_i to a lower-energy state E_f , energy in the form of electromagnetic radiation is given off and the frequency of the radiation ν_{if} is related to the energy difference between the two states by the equation:

$$\nu_{if} = \frac{E_i - E_f}{h} \quad (12)$$

h is the well-known Planck constant of action and has a value of 6.625×10^{-27} erg sec. This frequency and others which connect discrete energy states consistent with the so-called selection rules for the atom can be pictured as a set of frequencies characteristic of the particular atom. By analogy with resonance in a mechanical oscillator absorption of energy at a maximum rate might be expected to occur when there is radiation incident on the atom with a frequency which matches one of these frequencies. In quantum mechanics, the meaningful index for such a process taking place is the probability that an energy transition will occur a time t after being exposed to the radiation. When calculated, this probability is found to be proportional to

$$\frac{1 - \cos 2\pi(\nu_{ij} - \nu)t}{(\nu_{ij} - \nu)^2} \quad (13)$$

(see references 3 and 4).

Even though this probability function is not of the same mathematical form as the expression for rate of energy absorption derived in the section on "Resonance in a Simple Mechanical Oscillator," because of its dependence on $(\nu_{ij} - \nu)^2$ instead of $(\nu_0^2 - \nu^2)$, it can be looked upon as establishing a condition for a resonance type process. It predicts that the likelihood of resonance absorption becomes very large when the external frequency approaches something resembling a natural frequency of the atom. A distinctive example of this kind of resonance absorption is the phenomenon in the solar spectrum described in the introduction. The radiation from the hot core of solar gases is characterized by a continuous spectrum. In passing through the cooler gases of the sun's outer atmosphere, radiation will be absorbed at those frequencies which match frequencies of atomic transitions in the cool gas. These blocked out lines in the continuous spectrum show up as dark lines against the background. They are called Fraunhofer lines. The rare gas helium was first discovered as a consequence of these observations.

Another important class of atomic resonance phenomena and one in which there is great current interest concerns the behavior of nuclei and electrons in magnetic fields (see MAGNETIC RESONANCE and references 5, 6 and 7). If we start by using a classical analogy, the nucleus and electron can be thought of as being like little spinning magnetized tops with the axis connecting magnetic poles lying along the axis of spin. Being like a spinning top means the particle possesses the mechanical property of angular momentum. Being magnetized, the top will be subject to a torque if a magnetic field is applied along a direction different from that of the magnetic axis. Since the particle has angular momentum, the torque will cause the spin axis to precess about the direction of the applied magnetic field with a characteristic frequency called the Larmor frequency. This is similar to the precession of a spinning top due to the earth's gravity. The Larmor frequency is equal to $g\mu_0 H/h$, where H is

the applied field and h is Planck's constant. μ_0 is a nuclear magneton if the particle is a nucleus and a Bohr magneton if it is an electron. A magneton is the unit in which the particle's magnetic moment is measured. A nuclear magneton has a value of 5.050×10^{-24} erg/gauss. A Bohr magneton has a value of 0.927×10^{-20} erg/gauss. g is the Lande-factor, a special number in atomic and nuclear physics which appears in the ratio of the magnetic moment of a particle to its angular momentum. It is 2.00 for a free electron and 5.58 for a proton. If a radio-frequency (rf) magnetic field is applied in a plane perpendicular to H with a frequency exactly equal to the Larmor frequency, the particle will be subjected to an additional steady torque tending to change its angle with H . But here the quantum nature of the particle must affect the physical interpretation of what happens. The component of the angular momentum, and hence of the magnetic moment, along H can assume only certain discrete values. This means that the energies corresponding to these orientations can have only discrete values differing in an amount which turns out to be $g\mu_0 H$. If the frequency of the rf magnetic field is exactly $g\mu_0 H/h$, it can be demonstrated by a transition probability argument that transitions to either a higher or a lower magnetic energy state become very likely. If this is a system of particles in thermal equilibrium satisfying Boltzmann statistics, there will be more atoms in lower than higher states, and thus there will be a net absorption of energy from the rf field. This is called nuclear magnetic resonance (NMR) when changes in nuclear magnetic states are induced and electron paramagnetic resonance (EPR) when changes in electronic magnetic states are induced. The NMR frequency for a proton in water is $4.258H$ kc where H is in gauss, and the EPR frequency for a free electron is $2.80H$ Mc. The applications of magnetic resonance techniques to the studies of solids have proven very fruitful during the last decade. A very well-defined electron spin resonance can be observed if a sample of diphenylpicryl-hydrazyl (DPPH) is placed in a 3-cm waveguide and a transverse magnetic field of 3570 gauss applied. The resonance appears as a distinct reduction in the amount of microwave power transmitted past the region where the sample is located.

ROBERT LINDSAY

References

1. Feynman, R. P., Leighton, R. B., and Sands, M., "Lectures on Physics," Vol. 1, Ch. 23, Addison-Wesley Publishing Co., 1963.
2. Lindsay, R. B., "Physical Mechanics," Third edition, Ch. 10, Princeton, N. J., D. Van Nostrand Co., 1962.
3. Lindsay, R. B., "Influence of Environment on Transmission of Energy," *Am. J. Phys.*, **28**, 67-75 (1960).
4. Lindsay, R. B., and Margenau, H., "Foundations of Physics," Ch. 3, New York, John Wiley & Sons, 1936.

5. Pake, G. E., *Sci. Am.*, 199 58 (August, 1958).
6. Feynman, R. P., Leighton, R. B., and Sands, M., "Lectures on Physics," Vol. II, Chs. 34 and 35, Addison-Wesley Publishing Co., 1963.
7. Dekker, A. J., "Solid State Physics," Ch. 20, Englewood Cliffs, N. J., Prentice-Hall, Inc., 1957.

Cross-references: MAGNETIC RESONANCE, ELECTRON SPIN; DYNAMICS; SOLAR PHYSICS.

RHEOLOGY

When a force f is applied to a body, four things may happen. The body may be accelerated, strained, made to flow, and slid along another body. If these four responses are added to each other, one can write for motion in one direction

$$f = m\ddot{x} + r\dot{x} + sx + f_0 \quad (1)$$

where m is the mass, r is a damping parameter related to viscosity, s to elasticity, and f_0 to the yield value. Evaluation of the coefficients, m , r , and s involves the measurement of displacements, x , and their time derivatives in a manner which links these kinematic variables via an equation of state such as Eq. (1) to stress, σ (force per unit area), and its time derivatives.

Scope of Rheology. In contrast to the discipline of mechanics wherein the responses of bodies to unbalanced forces are of concern, rheology concerns balanced forces which do not change the center of gravity of the body. Rheology is the science of deformation and flow, and therefore is concerned primarily with the valuation of the coefficients r and s of Eq. (1). The coefficients account for most of the energy dissipated and stored, respectively, during the process of distorting a body.

Most rheological systems lie between the two extremes of ideality: the Hookean solid and the Newtonian liquid.

Measurements of Viscosity and Elasticity in Shear. Simple Shear. Shear viscosity η and shear elasticity G are determined by evaluating the coefficients of the variables \dot{x} and x , respectively, which result when the geometry of the system has been taken into account. The resulting equation of state balances stress against shear rate $\dot{\gamma}$ (reciprocal seconds) and shear strain γ (dimensionless) as the kinematic variables. For a purely elastic, or Hookean, response,

$$\sigma = G\gamma \quad (2)$$

and for a purely viscous, or Newtonian, response

$$\sigma = \eta\dot{\gamma} \quad (3)$$

As a consequence, G can be measured from stress-strain measurements, and η from stress-shear rate measurements.

Elasticoviscous behavior is described in terms of the additivity of shear rates:

$$\dot{\gamma} = \frac{\sigma}{\eta} + \frac{\dot{\sigma}}{G} \quad (4)$$

whereas viscoelastic behavior is characterized by the additivity of stress according to Eq. (1):

$$\sigma = G\gamma + \eta\dot{\gamma} \quad (5)$$

See reference 1 for further information on rheological bodies.

Relaxation. Numerous attempts have been made to fit simplified mechanical models to the two behavior patterns described by Eqs. (4) and (5). One can picture the elastic element as a spring arrayed in a network parallel with the viscous element to give essentially a (Kelvin) solid with retarded elastic behavior, wherein

$$\frac{\eta_k}{G_k} = \lambda \text{ sec (retardation time)} \quad (7)$$

or as a series (Maxwell) network which flows when stressed or relaxes under constant strain:

$$\frac{\eta_m}{G_m} = \tau \text{ sec (relaxation time)} \quad (8)$$

and transient experiments may be designed to measure these parameters singly. In real systems, a single relaxation (or retardation) time fails to account for experimental results. A distribution of relaxation times exists (see RELAXATION).

Dynamic Studies. When Eq. (1) is written in the form

$$\ddot{x} + 2k\dot{x} + \omega_1^2 x = 0 \quad (1b)$$

the equation suggests that the variation in stress be cyclic. Rheometers are designed so that the system may oscillate in free vibration of natural resonant frequency ω_1 , or else so that a cyclic shearing stress of the form $f_0 \cos \omega t$ is impressed on the sample over a frequency range which spans ω_1 . In neither case is the material strained beyond its range of linearity. Equation (1b) represents a damped harmonic oscillator, providing that the coefficients are constant (i.e., providing that they do not depend on the strain magnitude). Not all systems meet this requirement in the strict sense, with the result that one of the first consistency checks which the experimenter makes is for linearity. Doubling the amplitude of oscillation should double the stress and should not change the phase relationships between the cyclic stress and the deformation.

See RESONANCE and see reference 2 for more information on dynamic studies.

Time-temperature Equivalence. Steady-state Phenomena. The creep of a viscoelastic body or the stress relaxation of an elasticoviscous one are employed in the evaluation of η and G . In such studies, the long-time behavior of a material at low temperatures resembles the short-time response at high temperatures. A means of superposing data over a wide range of temperatures has resulted which permits the mechanical behavior of viscoelastic materials to be expressed as a master curve over a reduced time scale as large as twenty decades, or powers of ten (see references 3 through 5).

Polymeric materials generally display large G values (10^{10} dynes/cm² or greater) at low temperatures or at short times of measurement. As either of these variables is increased, the modulus drops, slowly at first, then attaining a steady rate of roughly one decade drop per decade increase in time. If the material possesses a yield value, this steady drop is arrested at a level of G which ranges from 10^7 downward.

Dynamic Behavior. The application of sinusoidal stress to a body leads inevitably to the complex modulus G^* , where

$$G^* = G' + iG'' = G' + i\omega\eta' \quad (9)$$

where G' is the in-phase modulus (σ/γ) which represents the stored energy, and G'' is the out-of-phase modulus ($\sigma/\dot{\gamma}$) representing dissipated energy (as its relation to η' suggests); the variable against which G' and η' are determined is the circular frequency ω . Superposition of variable temperature data or variable frequency data provides a master curve of the type described for the steady-state parameters.

Problems in Three Dimensions. State of Stress. The forces and stresses applied to a body may be resolved into three vectors, one normal to an arbitrarily selected element of area and two tangential. For the xy plane the stress vectors are σ_{xx} , and σ_{xy} , σ_{yx} , respectively. Six analogous stresses exist for the other orthogonal orientations, giving a total of nine quantities, of which three exist as commutative pairs ($\sigma_{xy} = \sigma_{yx}$). The state of stress, therefore, is defined by three tensile or normal components (σ_{xx} , σ_{yy} , σ_{zz}) and three shear or tangential components (σ_{xy} , σ_{yz} , σ_{zx}). The shear components are most readily applicable to the determination of η and G .

Strain Components. For each stress component σ there exists a corresponding strain component γ . Even for an ideally elastic body, however, a pure tension does not produce a pure γ_{xx} strain; γ components exist which constrict the body in the y and z directions.

The complete stress-strain relation requires the six σ 's to be written in terms of the six γ components. The result is a 6×6 matrix with 36 coefficients, k_{rs} , in place of the single constant. Twenty-one of these coefficients (the diagonal elements and half of the cross elements) are needed to express the deformation of a completely anisotropic material. Only three are necessary for a cubic crystal, and two for an amorphous isotropic body. These parameters are discussed in reference 6.

Similar considerations prevail for viscous flow, in which the kinematic variable is $\dot{\gamma}$.

RAYMOND R. MYERS

References

1. Reiner, Marcus, "Deformation, Strain and Flow," Second edition, London, H. K. Lewis & Co., 1960; "Lectures in Theoretical Rheology," Third edition, New York, Interscience Publishers, 1960.
2. Eirich, F. R. Ed., "Rheology, Theory and Applications," Three volumes, New York, Academic Press, 1956-1960.
3. Myers, R., Ed., "Transactions of the Society of Rheology," published annually 1957-1964 and semiannually 1964-.
4. Ferry, J. D., "Viscoelastic Properties of Polymers," New York, John Wiley & Sons, 1961.
5. Tobolsky, A. V., "Properties and Structures of Polymers," New York, John Wiley & Sons, 1960.
6. Alfrey, Turner, Jr., "Mechanical Behavior of High Polymers," New York, Interscience Publishers, 1948.

Cross-references: ELASTICITY, FLUID DYNAMICS, VISCOSITY.

ROTATION—CIRCULAR MOTION

Circular or rotary motion is very common and is best exemplified by a wheel rotating on an axle or a particle revolving in a circle about an attracting center. The case of a fixed axis will be considered in this section. This and related topics are elaborated on in greater detail in most college physics texts; in particular, "University Physics" by F. W. Sears and M. W. Zemansky is recommended.

Consider, as in Fig. 1, a line fixed in the plane of the wheel or orbit and terminating at the axis of rotation. A second radial line is fixed in the body and also terminating on the axis. The angle between these two lines specifies the position or orientation of the orbiting particle P or a section of the rotating wheel. Certain equations are more simply written if this angle is expressed in radians, but revolutions or degrees are sometimes used.

If θ changes from θ_1 to θ_2 as time increases from t_1 to t_2 , then the average angular velocity is defined as

$$\omega_{ave} = \frac{\theta_2 - \theta_1}{t_2 - t_1} = \frac{\Delta\theta}{\Delta t}$$

The instantaneous angular velocity, ω , is the value which this ratio approaches as $\Delta\theta$ and Δt become small. It is most advantageously expressed in radians per second.

Again, if ω is not constant but changes from ω_1 to ω_2 as t increases from t_1 to t_2 , then the

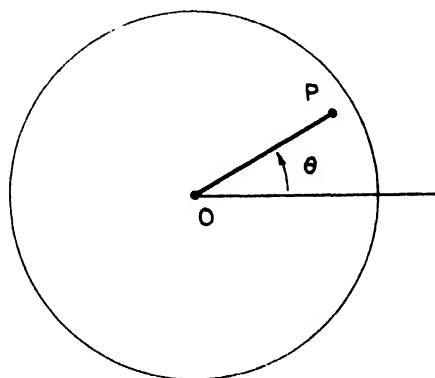


FIG. 1. Body rotating about a fixed axis through O.

average angular acceleration is

$$\alpha_{\text{ave}} = \frac{\omega_2 - \omega_1}{t_2 - t_1} = \frac{\Delta\omega}{\Delta t}.$$

The instantaneous acceleration, α , is the limit which this ratio approaches as $\Delta\omega$ and Δt become small; α is most advantageously expressed in radians per second per second.

In calculus notation $\omega = d\theta/dt$ and $\alpha = d\omega/dt$. Inversely,

$$\theta = \int \omega dt$$

and if ω is known as a function of t , then the integration may be effected.

Also

$$\omega = \int \alpha dt$$

and may again be evaluated. A simple situation is a constant α , and this case is discussed fully in the reference cited.

Next, consider a particle in a circular orbit of radius r , or a particle in the plane of a wheel a distance r from the center but not necessarily on the rim. Such a particle moves along a circular arc of length $s = r\theta$, where θ must be in radians; s and r are both linear quantities and have the same units of length: feet, meters, etc. Here linear implies a length, not necessarily along a straight line, and is contrasted with an angular quantity which implies an angle of some sort. The equation $s = r\theta$ is the connection between linear and angular quantities and is the basis for a useful analogy between motion along a straight line and rotation about a fixed axis.

For this same particle moving in a circular arc, if s and θ change by Δs and $\Delta\theta$ and in the time Δt , then

$$\frac{\Delta s}{\Delta t} = r \frac{\Delta\theta}{\Delta t} \text{ or } v = r\omega.$$

By the same reasoning, $a = \alpha r$. Of course v is the instantaneous velocity along the arc (tangential to the arc) and a is the instantaneous tangential acceleration.

There is another component of acceleration involved in this circular motion. The linear velocity v is a vector quantity. Whenever v changes, in either magnitude or direction, there is an acceleration. The acceleration of the preceding paragraph, $a = \alpha r$, arises because the particle is changing the magnitude of its tangential velocity, that is the particle is either speeding up or slowing down. But of course v , being along the tangent, is always changing in direction, and this amounts to an acceleration too. Derivations of this acceleration can be found in various texts and it turns out to be $a = v^2/r = \omega^2 r$. This acceleration is directed inward along the radius (it is thus at right angles to the tangential acceleration) and is sometimes called radial acceleration or centripetal acceleration. (A more advanced treatment of orbital motion would employ the polar coordinates (r, θ) . Then if the particle moves along the radius, there would be still another acceleration along the

radius, and the total radial acceleration in this notation is

$$\frac{d^2 r}{dt^2} = \omega^2 r$$

The concept of kinetic energy is first developed in translational motion and is defined as $\frac{1}{2}mv^2$ where m is the mass of some particle moving with the velocity v . Kinetic energy is a scalar quantity and if there are two particles, their kinetic energies just add. Applying this to rotation, a rigid body rotating about a fixed axis, may be broken into small particles of mass m_1, m_2, \dots having velocities v_1, v_2, \dots respectively. The total kinetic energy is then

$$\frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2 \dots$$

But $v_1 = \omega r_1$, $v_2 = \omega r_2$ where the ω is the same for all particles. So the kinetic energy equals

$$\frac{1}{2}m_1\omega^2r_1^2 + \frac{1}{2}m_2\omega^2r_2^2 \dots = \frac{1}{2}(\omega^2)(m_1r_1^2 + m_2r_2^2 + \dots)\omega^2.$$

The quantity in the parentheses, usually denoted by I , is called the moment of inertia of the rotating body, and it plays the role for rotation that m , the inertia, plays for translation. Thus kinetic energy $= \frac{1}{2}I\omega^2$ in analogy to kinetic energy $= \frac{1}{2}mv^2$. The moments of inertia of various bodies are listed in tables.

Extending this analogy, linear momentum is defined as mv . This suggests that angular or rotational momentum could be defined as $I\omega$. This is a satisfactory definition of angular momentum for a rigid body, but a more general definition of angular momentum is $mvr \sin \theta$ where θ is the angle between v and r . Angular momentum is a very fundamental quantity because it remains constant for an isolated system. Also it is found to consist of integral multiples of a certain smallest amount of angular momentum. That is, angular momentum can be added to a body or taken from a body only in integral multiples of this smallest constant quantity. A photon for instance is found to possess a quantum of angular momentum. This property of angular momentum is particularly important for small bodies such as rotating molecules.

Now these accelerations, which in turn imply changes in rotational kinetic energy and angular momentum, are brought about by the application of forces. Let us consider the two cases for the two components of the accelerations, the radial and the tangential components. We of course will use Newton's law $F = ma$ where F is the resultant force applied to m , a is the acceleration of m , and a is in the direction of F . Thus a radial acceleration is associated with a radial force and a tangential acceleration with a tangential force.

Case I. Force in the Radial Direction. Since there is no tangential force, the tangential acceleration is zero and the particle (or wheel) moves with constant angular velocity and constant tangential velocity. Further, for strictly circular motion, the acceleration is just $\omega^2 r$ directed toward the center and from Newton's second law, $F = m\omega^2 r$ also directed toward the

center. This force is called the centripetal force and is the radial force which must be applied to a mass m in order to keep it moving in a circular path. If this force were not applied, the particle would not move in the circular path, but it would move along a straight line in accordance with Newton's first law.

There are many examples of this centripetal force. If a mass tied to a string is whirled in a circle, then this force must be applied to the mass by the string. If a car is rounding a curve, then the highway must exert on the car a force which is toward the center of the path. In either example, if the force is not exerted (that is, there is no force on the mass) the particle would move along a straight line rather than in a circle.

Case II. Force in the Tangential Direction. A particle speeding up or slowing down in its circular path has a tangential acceleration, a , which is related to its angular acceleration by $a = r\alpha$. Again by Newton's second law this requires a tangential force $F = ma = mr\alpha$. This can be extended to the case of many particles, as in a rotating wheel, by the following maneuvers. Multiply both sides by r obtaining $rF = mr^2\alpha$. The left hand side is recognized as the torque about the axis of rotation of the force F . Of course there is one such equation for each of the many particles in the wheel. Adding all such equations one gets on the left hand side just the total torque, usually denoted by L . On the right hand side, since α is the same for all particles, the summation will introduce a combination we have seen before and the sum is just $I\alpha$. So $L = I\alpha$ which is the rotational analog of $F = ma$. Thus torque, L , is the analog of F and, as before, I is the analog of m .

Centrifugal force is referred to in many text books and Coriolis force, while much less common, is sometimes mentioned. In contrast, some texts scrupulously avoid these two forces or show a greater distaste for centrifugal force than CORIOLIS FORCE. An attempt will be made to reconcile these points of view. A key idea is the fact that different observers watching the same chain of events may record quite different observations and hence give quite different explanations. An inertial observer, that is, one at rest with respect to the fixed stars, and a rotating observer may do just that. Their explanations are equally valid, at least, the General Theory of Relativity admits both inertial and non-inertial observers on equal terms.

Consider first a ball tied to a string and whirled in a horizontal circle and then released. Any bystander is essentially an inertial observer and would see the ball fly off along the tangent in a straight line. Consider next the apparatus shown in Fig. 2. It consists of a horizontal turntable, T , on which there is a steel ball, B , and an electromagnet, M . The ball rolls on a toy "magic slate," S , on which it leaves a trace of its path. The slate plays the role of a rotating observer and the trace left on it is a record of how it observes the motion. The procedure is to energize the magnet with the ball in contact with it, set the table into rotation

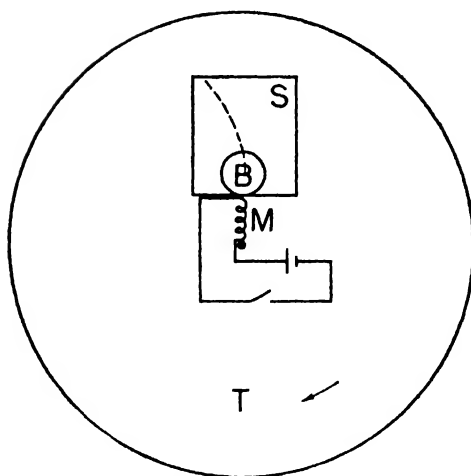


FIG. 2. Reprinted from *American Journal of Physics*, 27, (6), 429 (September, 1959).

and then release the ball. The dashed line represents the path followed. Notice that on the slate the ball starts rolling in a radial direction and then curves. The inertial observer sees a tangential motion of a released ball, while the rotating observer sees a radial motion.

Before the ball is released, the inertial observer says there is just one force acting on the ball, the tension in the string. This force on the ball is directed toward the center of the circle and is called centripetal force. The inertial observer also sees the ball accelerating toward the center and says that the centripetal force produces the centripetal acceleration. When the ball is released the centripetal force disappears, there is no force on the ball and the ball then travels in a straight line at constant speed.

In contrast, the rotating observer says that before the ball is released it is at rest. It has no motion and no acceleration with respect to the slate and is therefore in equilibrium. Hence two equal but opposite forces act on the ball. These two forces are a centripetal force applied by the magnet and a centrifugal force which acts radially outward. When the ball is released, the centripetal force ceases. This allows the centrifugal force, which continues to act, to accelerate the ball radially outward. Then follows a curvature of the path caused, according to the rotating observer, by the Coriolis force. This force appears when the ball acquires a velocity, v , with respect to the rotating observer and turns out to be always perpendicular to v . It is only the rotating observer who uses the concepts of centrifugal and Coriolis force. Notice also that while centrifugal and centripetal force are equal and opposite, they are not the action and reaction of Newton's Third Law as is implied in some texts. Actually since they act on the same body, they are equilibrants one of the other.

ARTHUR G. ROUSE

Cross-references: CENTRIFUGE, DYNAMICS.

S

SCHRÖDINGER EQUATION

The Schrödinger equation, first obtained in 1926¹, was an extension of de Broglie's hypothesis, proposed two years earlier,² that each material particle has associated with it a wavelength λ related to the linear momentum p of the particle by the equation

$$\lambda = \frac{h}{p} \quad (1)$$

where h is Planck's constant. Since any sinusoidally varying wave motion of amplitude ψ and wavelength λ satisfies the differential equation*

$$\nabla^2\psi + \frac{4\pi^2}{\lambda^2}\psi = 0 \quad (2)$$

matter waves would obey the equation

$$\nabla^2\psi + \frac{4\pi^2}{h^2}p^2\psi = 0 \quad (3)$$

In particular, a particle of mass m with no forces acting on it has energy $E = \frac{1}{2}mv^2 = p^2/2m$, or $p^2 = 2mE$. Equation (3) may thus be written

$$\nabla^2\psi + \frac{8\pi^2m}{h^2}E\psi = 0 \quad (4)$$

This is the Schrödinger equation for a free particle. Of greater importance is the equation for a bound particle for which the binding force can be related to a potential energy V . In this case, the total energy of the particle is equal to the sum of the kinetic and potential energies, i.e., $E = p^2/2m + V$, or $p^2 = 2m(E - V)$. Equation (3) thus becomes

$$\nabla^2\psi + \frac{8\pi^2m}{h^2}(E - V)\psi = 0 \quad (5)$$

This is known as the time-independent Schrödinger equation. Its great utility lies in the fact that it enables one to calculate energy levels, or eigenvalues, of the energy E (see QUANTUM THEORY). The remarkable agreement of the results of the equation with experimental fact has led to its being regarded as one of the fundamental equations of physics.

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

The more advanced formulation of the Schrödinger equation is based on the operator concept of quantum theory together with the Hamiltonian methods of classical mechanics. Specifically, to each dynamical variable q_i , there is a conjugate momentum p_i . The fundamental postulate of quantum mechanics states that for each pair (q_i, p_i) , the equation

$$p_i\psi = -i\hbar \frac{\partial\psi}{\partial q_i} \quad (6)$$

holds, where $\hbar = h/2\pi$. That is, $i\hbar\partial/\partial q_i$ is an operator, which, when applied to the wave function ψ , is equivalent to multiplying by p_i . In the Hamiltonian function $H(p_i, q_i)$, each p_i is replaced by its corresponding operator. The result is the Hamiltonian operator $H(-i\hbar\partial/\partial q_i, q_i)$, which, when it operates on the wave function, is equivalent to multiplying by the energy E , viz.,

$$H(-i\hbar\partial/\partial q_i, q_i)\psi = E\psi \quad (7)$$

This is the generalized form of the time-independent Schrödinger equation. It reduces to Eq. (5) when applied to a single particle, but it has the advantage that the application to systems of many particles can be carried out in a straightforward manner.

The time-dependent form of the Schrödinger equation is obtained by replacing E by its operator $E \rightarrow i\hbar\partial/\partial t$, so that

$$H\left(-i\hbar \frac{\partial}{\partial q_i}, q_i\right)\psi = i\hbar \frac{\partial}{\partial t}\psi \quad (8)$$

For applications of the various forms of the Schrödinger equation, the reader should consult the article on WAVE MECHANICS and references therein.

GRANT R. FOWLES

References

1. *Ann. Physik*, **79**, 361, 489 (1926).
2. *Phil. Mag.*, **47**, 446 (1924).

Cross-references: MATRIX MECHANICS, QUANTUM THEORY, WAVE MECHANICS, WAVE MOTION.

SECONDARY EMISSION

In its most general sense, secondary electron emission refers to the ejection of electrons from matter under the impact of rapidly moving particles such as electrons, ions, or neutral atoms. When the atoms of the material bombarded are in the gaseous state, the phenomenon is more commonly referred to as impact ionization. In practice, the term is mainly used in connection with the emission of electrons from solids under electron bombardment as in photomultipliers, and for the ejection of electrons at the cathode of gas discharge devices under the action of positive ions.

Historical Background. Secondary emission from solids was first observed with electrons as incident particles. The effect was discovered by Austin and Starke in 1902 in the course of studies on the reflection of fast electrons from metals. Upon increasing the angle of incidence of the primary electrons away from the normal, Austin and Starke observed that the current into the target first decreased to zero and then actually reversed. They concluded that not only are the incident electrons reflected but other electrons are also ejected from the plate such that the total number leaving exceeds the number arriving. The effect was found to be strictly proportional to the number of primary electrons. Furthermore, it was found to be more pronounced for metals of high density, and it was observed to increase rapidly with the angle of incidence. Shortly thereafter, P. Lenard (1903) showed that the secondaries are emitted diffusely, that they possess low energies of the order of a few volts independent of primary energy or target material, that the yield goes through a maximum at a few hundred volts, and that the process occurs in insulators as well as metals. Lenard concluded that the formation process is fundamentally an atomic one and therefore largely independent of the state of aggregation. This deduction has been confirmed in recent years.

The emission of secondary electrons from solids under ion bombardment was first demonstrated in 1905 by J. J. Thomson and established to be very similar to the case of electron bombardment by C. Füchtbauer (1906). The secondaries were again found to have energies of a few volts, essentially independent of the nature of the bombarding ion, its velocity, or the properties of the solid bombarded. Shortly thereafter, N. R. Campbell (1911) established the same characteristics for the electrons emitted under α -particle bombardment, the so-called δ -rays.

Secondary Emission Under Electron Bombardment. The basic features of the phenomenon as clarified by subsequent investigations may be summarized as follows. For all pure metals and elemental semiconductors free from surface oxides or gas contamination, the total yield δ , defined as the total number of electrons emitted per incident primary electron, never exceeds a value of about 2. Out of this total, the number of backscattered electrons per incident primary η

may be as high as 0.5 for high atomic number elements and energies in excess of a few kilovolts. The distinction between the two kinds of emitted electrons is based on their energy distribution. The true secondaries are characterized by their low energies, peaked at about 2 eV for all metals, and a Maxwellian-like energy distribution with a mean value close to 5 eV independent of primary energy. The backscattered electrons above about 50 eV emerge with energies all the way to the primary energy, but only a few per cent are truly elastically reflected, having lost no appreciable energy in the solid (as indicated in Fig. 1).

The backscattered fraction η is almost constant with changing primary energy for low atomic number elements ($Z \lesssim 30$), varying nearly linearly with Z from 0.04 for Be ($Z = 4$) to 0.28 for Ni ($Z = 28$). For heavier elements, η increases slowly with primary energy from values less than for Ni to limiting values as high as 0.45 for Pt ($Z = 78$) above 5 kV. The yield of low energy secondaries $\Delta = \delta - \eta$ starts from zero at a finite voltage in the neighborhood of 10 to 15 eV and rises to a maximum value in the range of 0.5 to 1.5 between 100 and 700 volts for all metals. Thereafter, it decreases steadily until at energies above a few kilovolts, Δ becomes less than the backscattered fraction η .

In sharp contrast to thermionic or photoelectric emission, neither the onset of emission nor the maximum yield is directly related to the work function of the metal. Instead, it has been found that the maximum yield Δ_m increases steadily within each period of the atomic table as successive elements are added; the alkali metals have the lowest yield in each case as illustrated for the fourth period by Fig. 2.

The presence of surface impurities or oxides can increase the yield considerably, primarily owing to the fact that secondaries can travel larger distances in insulators than in metals, where they lose energy rapidly to the conduction electrons. Measurement of the yield for insulators, such as BaO, MgF₂, MgO, and KCl, have shown that yields of 6 to 10 or higher can be obtained for such materials depending on the method of preparation, the yields remaining well above unity even at primary energies of many kilovolts. It is for this reason that in technical applications in which it is desired to obtain the largest possible yields, one utilizes alloys which upon heat treatment form an oxide layer containing BeO or MgO on their surfaces. The most common of them are alloys of Cu-Be and Ag-Mg.

In addition to the larger mean free path for secondaries, such complex insulating surfaces often show an additional enhancement of their yield owing to internal electric fields, which act to increase the fraction of electrons able to escape, thus resulting in yields often many times those normally attainable. The presence of charging effects is characterized by a strong increase in the relative number of very low-energy secondaries. It is generally accompanied by a time delay in the emission process, which is long compared

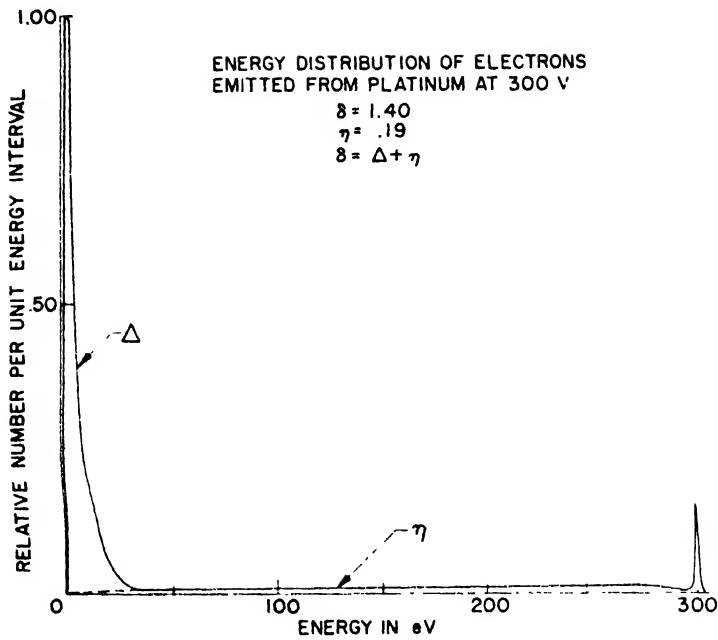


FIG. 1. Typical energy distribution of the electrons emitted from metals under electron bombardment by 300-v primaries; δ is the total yield, Δ the yield of true secondaries, and η the backscattered fraction.

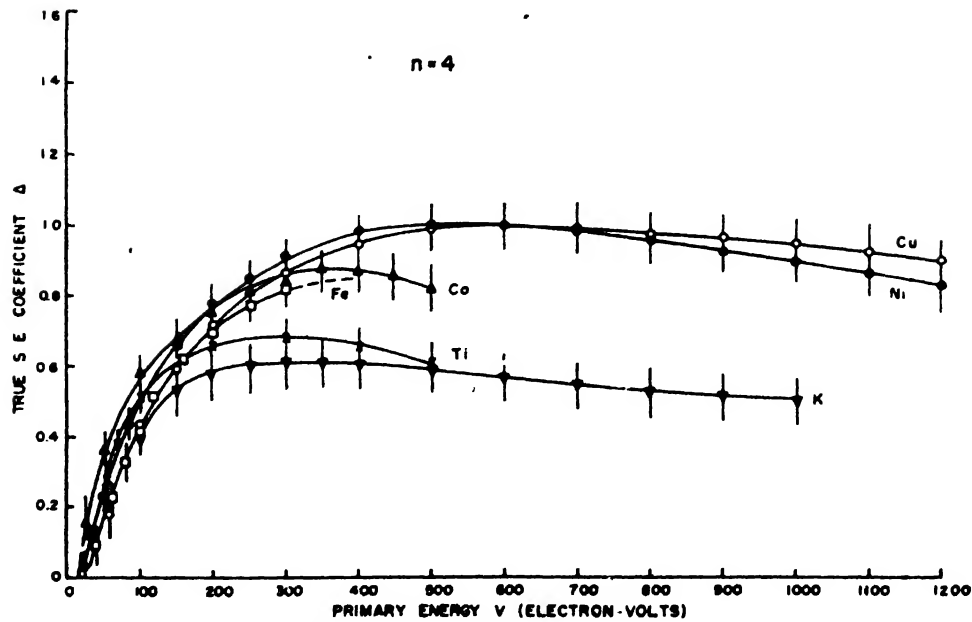


FIG. 2. Typical secondary-electron yield curves for metals. Data shown for elements in the 4th period of the atomic table.

with that for normal secondary emission, which is completed in less than 10^{-12} second.

Secondary Emission Under Ion Bombardment. As already noted, secondary emission under ion bombardment resembles closely the phenomenon under electron bombardment, especially for very high ion energies, when the velocities of the incident ions exceed the orbital velocities of electrons in the atoms of the solid. The principal difference is the somewhat larger yield for ions, brought about by their greater rate of energy loss and therefore ionization near the surface. For protons, the maximum yield is close to 4 occurring at an incident energy of about 100 kV. For α -particles or helium ions, the maximum yield is almost four times as large since the rate of ionization is proportional to the square of the effective charge. Another important difference in the case of fast ions is that all clean metals show closely the same yield at a given energy, because of a proportionality between the probability of secondary formation and absorption in all metals. Again, the presence of oxide layers can greatly enhance the attainable yield. For ions of very low velocity, the phenomenon becomes more complex since the relative values of the ionization potential and the work function of the solid begin to play a dominant role. The electron-emission process is then determined primarily by the energy made available when an electron from the metal drops into the ionized outer level of the approaching ion, rather than by its kinetic energy and state of charge.

Theory. In contrast to thermionic, photoelectric, and field emission, secondary emission under the impact of high-speed particles is an energetic process for which the usual free-electron model of a metal becomes inadequate. Furthermore, in the case of electrons falling on a solid, the strong inelastic scattering by the large number of more firmly bound electrons is the principal factor determining the depth at which the secondaries are formed. The theoretical description therefore consists of finding an expression for the ionization density as a function of depth, and then calculating the probability that the secondaries formed will reach the surface. When the primary particle velocity is high, the theory of Bohr and Bethe for the stopping of charged particles in gases may be used to calculate the scattering and ionization of the incident particles. In the special case of high-energy ions, scattering is negligible and a simple theory was formulated by E. J. Sternglass (1957) that explains the maximum in the yield as reflecting the maximum in the ionization probability for the atoms of the solid.

For the case of low-energy electrons incident on the surface, it is possible to arrive at an approximate expression for the mean depth at which the electrons have been completely scattered and where most of the secondaries are formed. In view of the known rapid exponential absorption of the secondaries in metals, with a mean-free path of only a few atomic layers and an energy expenditure per secondary formed similar to that for gases (~ 30 eV), good agreement with the

observed yields was obtained by E. J. Sternglass (1951). The shape of the yield curve is found to reflect the opposing effects of an increase in the number of secondaries formed and the decrease in their chance of reaching the surface as the primary energy is increased. In insulators, the escape probability is greater, thus leading to larger yields at high primary energies.

For the case of low-velocity rare-gas ions incident, the band-model of a metal can be used to calculate the energy available for the ejection of a conduction electron when a metallic electron neutralizes the incoming ion. The observed yield and energy distribution of the secondaries can then be accounted for rather well as shown by H. D. Hagstrum (1954). At the present time, the principal theoretical problem remaining is to establish the precise role of the surface potential barrier in determining the fraction of the secondaries able to escape.

Applications. The principal application of secondary emission in the field of electronics has been in the amplification of weak electrical currents first suggested by J. Slepian in 1917. The most common use is in the intensification of photoelectrons by successive steps of multiplication in photomultiplier tubes, over-all gains of 10^7 being readily achieved at a very small loss in signal-to-noise ratio. The phenomenon also plays an important role in many types of television camera tubes such as the iconoscope and the image orthicon, and more recently in image-intensifying tubes based on secondary emission from a series of thin foils. Other important applications exist in storage tubes, magnetrons, and high-gain multiplier receiving tubes. Secondary emission under ion bombardment is a fundamental process involved in most gas-discharge devices, and it is also believed to play an important role in high-voltage breakdown phenomena.

E. J. STERNGLASS

References

- Bruining, H., "Physics and Applications of Secondary Electron Emission," New York, Pergamon Press, 1954.
- Kollath, R., "Sekundarelektronen-Emission Fester Körper bei Bestrahlung mit Elektronen", in "Handbuch der Physik," Vol. XXI, pp. 232-303, Berlin, Springer, 1956.
- Dekker, A. J., "Secondary Electron Emission," in Seitz, F., and Turnbull, D., Eds., "Solid State Physics," Vol. 6, pp. 251-311, New York, Academic Press, Inc., 1958.
- Hachenberg, O., and Brauer, W., "Secondary Electron Emission from Solids," *Advan. Electron.*, **11**, 413 (1959).

Cross-references: ELECTRON, PHOTOELECTRICITY, PHOTOMULTIPLIER TUBE, THERMIONICS.

SEISMOLOGY

Seismology, the study of earthquakes and attendant phenomena, provides the bulk of man's detailed knowledge of the earth's interior, largely as a result of investigations of the complex seismic waves which propagate throughout the earth following an earthquake. Analyses are based principally on ray theory for elastic body waves of the dilatational and shear types, and on wave theory for traveling elastic surface waves and their standing wave counterparts, the free oscillations of the earth. Seismological studies have, among other things, defined the configuration and structure of the earth's nonrigid outer core, the probably rigid inner core, the rigid mantle with its low-velocity layer at a depth of a few hundred kilometers, and the deeper regions of the earth's crust. In practice, the science includes widely diverse activities ranging from operation of unusual seismographs in the most remote locations of the earth to application of highly sophisticated mathematical techniques to problems involving complex earth models.

Excluding certain near-surface regions, the velocities of dilatational seismic waves range from about 5 km/sec in parts of the crust to a maximum of $13\frac{1}{2}$ km/sec at the base of the mantle; corresponding shear wave velocities range from 3 to 8 km/sec. The shortest periods of interest in the study of waves from distant earthquakes are of the order of $\frac{1}{3}$ second (frequency = 3 cps); the longest periods are about 53 minutes (frequency ≈ 1 cycle/hour), and they correspond to a free oscillation of the earth in the fundamental spheroidal mode.

Free oscillations of measurable amplitudes are generated only by the largest earthquakes. Body and surface waves from large nuclear explosions and many earthquakes may be detected throughout the world; seismic waves from chemical explosions are adequate for exploration to depths of about 50 km or less. Seismic techniques are used extensively in petroleum exploration and for various other geologic purposes.

The largest earthquakes probably release between 10^{24} and 10^{25} ergs in the form of seismic waves, and the few largest shocks account for most of the energy released in this form. Mean annual release is estimated at 9×10^{24} ergs.

Earthquake size as determined by instruments is measured on a logarithmic scale called the Richter *magnitude* scale. In one variation of this scale, the very largest shocks have magnitudes slightly greater than $8\frac{1}{2}$. Energy in ergs is given empirically by $\log E = 11.4 + 1.5M$, where M is the magnitude. A noninstrumental scale for measuring earthquake effects at a given location is called the *intensity* scale; the popular Modified Mercalli version has divisions from I to XII. Whereas an earthquake may produce a range of intensities, it is in principle characterized by a single value of magnitude.

The frequency of occurrence of earthquakes increases by about a factor of 8 or 10 per unit of magnitude as the magnitude decreases. On the

average, only about 25 shocks with a magnitude of 7 or more occur each year, but it has been estimated that there are at least one million earthquakes per year, most of them quite small. About 5000 to 10,000 of these are routinely located and studied.

Most earthquakes occur in certain narrow world-circling belts separated by relatively stable blocks. The circum-Pacific belt is by far the most important, accounting for about 80 per cent of the total activity. Other important features are the trans-Eurasian belt and the mid-ocean belt. The foci, i.e., the points of initiation of the first seismic waves, of most shocks are shallow, i.e., at depths of 60 km or less, but some are as deep as 700 km. Most large shocks are followed by a series of smaller aftershocks which occur in the same general region as the main shock. Aftershocks generally decrease in size and in frequency of occurrence with time but may persist for more than a year following a very large shock. Some earthquakes are preceded by one or a few foreshocks.

The most popular model of the earthquake mechanism is based on the elastic rebound theory which calls for gradual accumulation of elastic strain in a region prior to the release of the strain energy through rupture or by slippage along a preexisting fault. This model explains many features of many earthquakes but is clearly oversimplified and may not be applicable at all to deep shocks. Precise prediction of earthquake occurrence in time and space is not now possible, but current and prospective research offer some promise.

In addition to transient seismic phenomena such as those due to earthquakes and explosions, there is a continuous background noise in the earth which is measurable with modern seismographs over a wide range of periods. For periods of about one second, ground amplitudes at very quiet locations are less than 1 millimicron. Between periods of 4 and 9 seconds, there is a sharp peak in the noise spectrum. The corresponding waves are called storm microseisms and are related, probably through ocean waves, to meteorological disturbances at sea.

Most modern seismographs are of the inertial type, depending upon measurement of the relative displacement between a point fixed to the earth and a mass loosely coupled to the earth. Other instruments measure relative displacement between two points in the earth. There are approximately one thousand seismograph stations in the world. Attempts are currently under way to operate seismographs on the floor of the deep ocean and on the moon.

JACK OLIVER

References

- Bullen, K. E., "Introduction to the Theory of Seismology," Third edition, Cambridge, Cambridge University Press, 1963.

- Coulomb, J., and Jobert, G., "The Physical Constitution of the Earth," New York, Hafner Publishing Co., 1963.
- Eiby, E. A., "About Earthquakes," New York, Harper & Bros., 1957.
- Ewing, M., Jardetzky, W. S., and Press, F., "Elastic Waves in Layered Media," New York, McGraw-Hill, 1957.
- Gutenberg, B., and Richter, C. F., "Seismicity of the Earth," Princeton, N. J., Princeton University Press, 1949.
- Hodgson, J., "Earthquakes and Earth Structure," Englewood Cliffs, N. J., Prentice-Hall, 1964.
- Jeffreys, H., "The Earth," Cambridge, Cambridge University Press, 1962.
- Richter, C. F., "Elementary Seismology," San Francisco, W. H. Freeman & Co., 1958.

Cross-references: GEOPHYSICS, VIBRATION, WAVE MOTION.

SEMICONDUCTORS

Semiconductors are materials whose electrical conductivity lies generally between that of metals (10^6 mho/cm) and insulators (10^{-10} to 10^{-20} mho/cm). However, the boundaries are indistinct, and on the one hand the properties of very impure semiconductors are quite similar to those of some metals (semi-metals such as bismuth, for example), and on the other hand, many semiconductors may be strongly insulating if highly purified or measured at low temperature.

As indicated by the above, the properties of semiconductors are extremely variable, and it is partly because of this that they are so interesting and so important both to the physics of the solid state and to electrical technology. The extreme sensitivity to impurities, to temperature, to incident light, to magnetic fields, and to nuclear radiation provides a rich field for the testing of theories of the solid state. Out of the same properties have also arisen such important technical contributions as transistors and integrated circuits, which have revolutionized electronics.

Among the elements, there are at least seven semiconductors. They are boron, silicon, germanium, gray tin, tellurium, selenium, and black phosphorus, while carbon (both diamond and graphite) has some semiconducting properties as do iodine, sulfur, and perhaps others.

There are numerous inorganic and some organic semiconducting compounds. The best known of the inorganics are the compounds of the third and fifth columns of the periodic table, such as gallium arsenide, indium arsenide, indium antimonide, and aluminum antimonide. Also important are the "2-6" compounds, of which cadmium sulfide, zinc sulfide, and zinc oxide are probably the best known. The ionic crystals such as the alkali halides (NaCl) and the alkaline earth oxides, may also be semiconductors under certain conditions. Among the organics, many dyes have electronic conductivity, and such materials as the phthalocyanines may be quite

conducting. Single crystals of members of the quinone family (TCNQ) have been prepared having resistivities as low as 1 ohm cm.

There is no uniformity in the appearance, physical properties, or crystal structure of semiconductors. Some are metallic-looking, while others may be practically transparent. Most are crystalline, but amorphous semiconductors (selenium) and liquid semiconductors (many of the molten sulfides and selenides as well as selenium) are also known. Unlike metals, semiconductors tend to be brittle and hard, rather than ductile. There is no uniformity in the method of preparation. Much of the effort in recent years has been devoted to methods of purification and single-crystal preparation. More than any other group of materials, semiconductors are sensitive to the perfection of the crystal lattice, as well as to the content of foreign impurities. Many semiconductors, especially silicon and germanium, can be drawn in large single crystals from the melt (Czochralski technique). Because of their importance as transistor materials, extensive research has made them probably the purest and most perfect materials available in science.

Other crystals may be grown from the vapor phase, prepared by direct combination, or frozen in tubes in a temperature gradient (Bridgman technique) so as to produce single crystals (see CRYSTALLIZATION).

Basic Theory of Semiconductors. There are several aspects of the theory of semiconductors, and all are needed to understand the electrical properties. The first is the theory of electronic band structure of solids, the second is the theory of electrical transport in solids (we deal here only with *electronic* as distinct from *ionic* conductivity).

The explanation of the temperature and impurity sensitivity of semiconductors in 1930 was one of the first triumphs of the quantum theory of solids, and the history of semiconductor research has been a parallel development of the quantum theory and of experimental research on materials. Even today, however, because of the complexity of the theory and of the calculations, the theory is still incomplete and approximate. One-electron theories still prevail, whereas many-electron theories are needed to treat the problem adequately. As more powerful computers are assigned to the calculations, we shall see rapid extension of theoretical knowledge particularly in terms of details of the properties of various materials.

The "bands" of semiconductor theory are energy bands which are derived from the atomic energy levels of, say, the individual germanium atom, broadened by the interactions between neighboring atoms in the crystal (see ENERGY LEVELS). The upper band, the conduction band, normally empty, corresponds to the first excited state, whereas the lower band, the so-called valence band, usually full of electrons, corresponds to the top normally occupied atomic level. Impurity atoms in the lattice may, at rather low

temperatures, give up an electron to the conduction band, and the semiconductor may then conduct electronically (*n*-type conduction). Other impurities, however, may *take up* electrons from the valence band, which then, rather remarkably, also contributes to conduction, the empty states in the band acting like *positive* charges with the same mass as the electron. Such semiconductors are *p*-type. At sufficiently high temperatures (room temperature for fairly pure germanium, 200°C for silicon), electrons from the valence band may be excited directly across the "forbidden gap" into the conduction band, and both the hole and the electron are conducting. Such semiconductors are *intrinsic* semiconductors, in contrast to impurity (extrinsic) semiconductors.

In the elementary sense, the energy-band picture showing the band gap of the material and the energy location of the levels from which the impurities deliver carriers can be considered a specification of the band structure. For many materials, however, we are now able to specify a more fundamental structure, including the "*E* vs *k*" curves and the "surfaces of constant energy" of the carriers. Only in recent years has this information become available for many materials. The *k* referred to is the coordinate of "*k* or momentum space", where *k* is the reciprocal wavelength of the electron considered as a wave

in the wave-mechanical sense. For a simple semiconductor, dealt with in older books and papers, the *E* vs *k* curve was a parabola, exactly as for an electron in free space, but with a mass different from that of an electron in free space to account for the effects of the lattice. The effective mass, determined by the curvature of the *E* vs *k* curve, may be greater or less than the free-electron mass. The surface of constant energy for a simple semiconductor is a sphere.

As properties of semiconductors became better known, it was realized that the simple model was not adequate. The parabolic *E* vs *k* curve for germanium's conduction band was replaced by a complicated set of curves having minima at *k* = 0 but also at *k* ≠ 0, and the constant energy surfaces, instead of being a single sphere, were sets of four ellipsoids, cigar-like in shape (20 times as long as thick) oriented along the (111) crystal directions (along the diagonals of the basic cube of the lattice). The *E* vs *k* curves for the valence band of Ge are nearly parabolic, but there are three of them. Two of them are centered at *k* = 0 and correspond to two different kinds of holes (light and heavy). They are examples of a *degenerate* valence band.

Once the basic band structure is known, STATISTICAL MECHANICS is available to determine the distribution of electrons over various possible states (conduction band, valence band, and

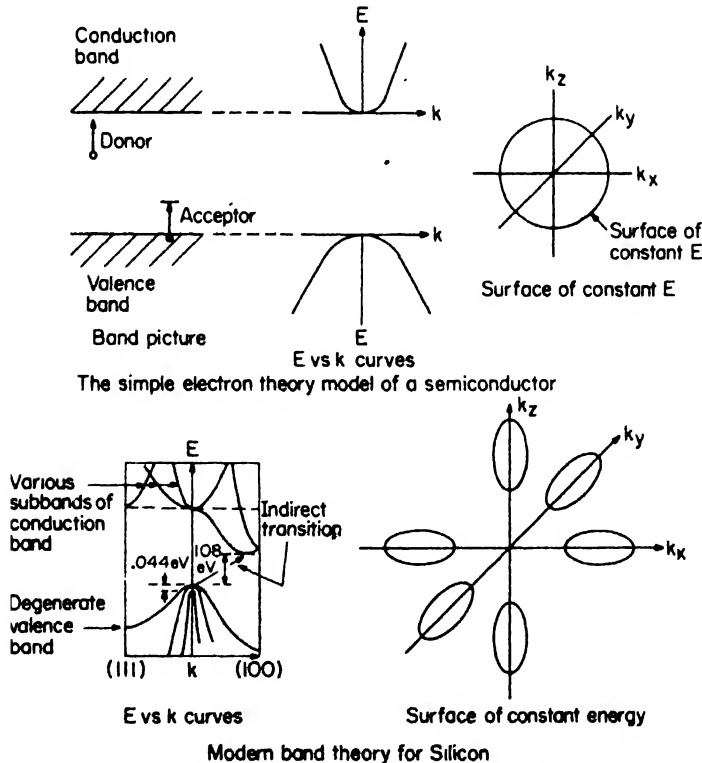


FIG. 1. Comparison of the older and newer pictures of the band structure of a typical semiconductor.

impurity levels) and to determine the carrier densities p and n . Another important task is to understand the *mobilities* of the carriers, μ_n and μ_p . Much knowledge of this has been gained on many materials during the past few years. The theory of mobilities requires understanding of the various *scattering* processes which hinder the motion of the carriers, the dependence on energy, and the influence of the scattering on the various experimental properties such as conductivity and Hall coefficient. Finally, knowing the carrier densities and mobilities, and one or two other parameters, theories are available to describe other experimental properties such as thermoelectric power, photoconductivity, optical absorption, magnetic susceptibility, and magneto-resistance.

The distribution function which determines the way electrons spread over the various states is the "Fermi" function $1/[1 + \exp(E - E_f/kT)]$, where E_f is the Fermi level, a characteristic marker on the energy-band picture for each sample. Since this function has a simple exponential form $\exp[-(E - E_f/kT)]$ for large E , there results an exponential variation of carrier density with temperature, one of the characteristic properties of semiconductors, the conductivity going to zero at absolute zero. For impurity semiconductors, *saturation* sets in at higher temperatures as the impurity supply of electrons becomes exhausted. The carrier density then remains constant until intrinsic conduction sets in when it again changes exponentially. For germanium, this exponential holds right up to the melting point, 938°C, where the conductivity value is about 1000 mho/cm. Liquid germanium is metallic, some ten times more conducting than the solid.

Determination of the scattering of carriers and calculation of experimental properties in terms of this scattering is equally as important as band structure studies. One must know the various mechanisms of scattering, their dependence on energy and temperature. Scattering of electrons or holes is usually due to (a) lattice scattering, (b) impurity scattering, or (c) scattering by imperfections such as vacancies or dislocations. Each of these has a different magnitude and dependence on energy. Lattice vibrations, or PHONONS, have a characteristic spectrum for each material, and much research in recent years has gone into determination of such spectra. Phonons may be of the *optical* or *acoustic* type (in optical vibrations, adjacent atoms move in opposite phases; in acoustic, the same phase); in some materials, the optical phonons may be the main source determining mobility; in others, the acoustic may be predominant. Many phenomena, such as thermal conductivity, optical absorption, radiative recombination, and tunneling have been used to gain information on phonons.

In some cases, for phenomena such as TUNNELING, phonons may be *required* to observe the phenomenon. Materials which show this effect are called *indirect* semiconductors, since the E vs k curve minima for the conduction band do not lie directly above the valence band maximum, and

hence momentum transfer, provided by phonons, is required for the transition. Germanium, silicon, and silicon carbide are examples of indirect semiconductors, while indium antimonide and gallium arsenide are the best known direct semiconductors. It is only in the latter group that semiconductor laser action is seen, because of the low efficiency of radiative recombination when phonons are required in the process.

Experimental Properties of Germanium. Let us briefly outline the experimental properties of semiconductors using germanium as an example.

Conductivity σ is of course a basic property. It depends upon carrier density and mobility according to the equation $\sigma = ne\mu$, where e is the charge on an electron. Conductivity in the simple theory and for cubic crystals is a scalar (pure number) quantity, but for anisotropic crystals it is a tensor, as are the effective masses, Hall coefficients, and others. The density of carriers is generally determined by use of the Hall coefficient, according to the equation $R = 1/ne$, in electromagnetic units. In modern work, R is given by r/ne , where r is the "Hall coefficient factor" determined by a ratio of two relaxation times for the electron distribution, averaged over the energy surfaces of the particular material. The value of r usually ranges from 0.5 to 2, and in older work was taken to be $3\pi/8$. For p -type semiconductors, the same equations prevail, with p instead of n .

Holes and electrons may coexist in equilibrium only in intrinsic semiconductors, while after light excitation or nuclear bombardment, the electron-hole pairs produced last only for a characteristic time called the *lifetime*. The conductivity for mixed conduction by holes and electrons is:

$$\sigma = e(n\mu_n + p\mu_p)$$

and the Hall coefficient

$$R = \frac{r}{e} \left(\frac{n\mu_n^2 - p\mu_p^2}{(n\mu_n + p\mu_p)^2} \right)$$

Germanium at room temperature has an intrinsic resistivity of 47 ohm cm, the carrier densities being each $2 \cdot 10^{13}/\text{cm}^3$. The electron mobility is $3800 \text{ cm}^2/\text{volt} \cdot \text{sec}^{-1}$; for holes, 1800. These values are much higher than for most metals and many semiconductors, which average 100 to $200 \text{ cm}^2/\text{volt} \cdot \text{sec}^{-1}$. Indium antimonide has the extraordinarily large mobility at room temperature of $70,000 \text{ cm}^2/\text{volt} \cdot \text{sec}^{-1}$. Such high-mobility materials are important for Hall-effect applications.

The sensitivity of germanium and other semiconductors to impurities is enormous. Even at room temperature, arsenic present to the extent of $2 \cdot 10^{12}$ atoms/ cm^3 changes the resistivity by 15 per cent. This is only one part in ten billion! Other donors besides arsenic are phosphorus, antimony, and bismuth, which give up their electrons to the conduction band with only 0.01 eV of energy required (ionization energy). The acceptors boron, aluminum, gallium, and indium take up electrons and have very nearly the same ionization energy.

TABLE 1. ROOM TEMPERATURE PROPERTIES OF SOME SEMICONDUCTORS

Material	Band Gap (eV)	Electron Mobility $\text{cm}^2/\text{volts} \cdot \text{sec}^{-1}$	Hole Mobility $(\text{cm}^2/\text{volts} \cdot \text{sec}^{-1})$	Dielectric Constant, ϵ	Lattice Constant (Å)	Density (g/cm ³)	Melting Point (°C)
Si	1.15	1900	480	11.8	5.42	2.4	1412
Ge	0.65	3800	1800	16.0	5.646	5.36	938
GaAs	1.35	8500	400	13.5	5.65	5.31	1280
GaSb	0.69	4000	650	15.2	6.095	5.62	728
InSb	0.17	70000	1000	16.8	6.48	5.775	525
SiC	3.0	60	8	10.2	4.35	3.21	2700
PbS	0.37	800	1000	17.9	7.5	7.61	1114
ZnO	3.2	190		8.5	(a) 3.24 (c) 5.18	5.60	1975
CdS	2.4	200		5.9	5.83	4.82	685
HgTe	0.2	22000	160		6.429	8.42	670

By now the effects of over 30 elements in germanium have been studied. Some such as gold may, under certain conditions, either donate an electron or accept as many as three electrons. Such an element is "amphoteric". Some elements such as tin or lead dissolve appreciably, but they seem to have no electrical action.

The boundary between the *p* and *n* regions of a single crystal is of great scientific and technical interest. A single "*p-n* junction" is the heart of the modern silicon or germanium rectifier, while two *p-n* junctions in close proximity (usually less than 10μ) form the basis for the *transistor*. The transition between the donor and acceptor impurities at the boundary requires a potential barrier to keep holes and electrons in place. When voltage is applied in one direction, however, large currents can flow (minority carrier injection), while with voltages in the opposite direction, no current can flow (or a very small one), and the barrier may withstand several thousand volts without breaking down. Transfer of injected carriers from a low-impedance emitter junction to a high-impedance collector junction biased with opposite voltage constitutes transistor action and gives power amplification. Transistors have also been made with gallium arsenide, indium antimonide, and other compounds.

Other Semiconductors. Table 1 gives the basic parameters of a number of semiconductors according to recently published data. For more details the reader should consult the references.

Most known semiconductors fit the general picture described above. There is a large class, however, about which we have only recently begun to get a clear understanding. These are the so-called narrow-band semiconductors, including most organic semiconductors as well as some inorganic ones. These materials may have a very high carrier density, but very low mobility which rises rapidly with temperature. The conductivity mechanism is thought to be a "hopping" mechanism from one atom to the next, rather than

electron wave motion perturbed by scattering centers.

W. C. DUNLAP

References

- Dunlap, W. C., "An Introduction to Semiconductors," New York, John Wiley & Sons, 1957.
- Smith, R. A., "Semiconductors," Cambridge, The University Press, 1959.
- Frederickse, H. P. R., "Properties of Semiconductors," in "American Institute Physics Handbook," Second edition, pp. 9-45 to 9-63, New York, McGraw-Hill Book Co., 1963.
- Shockley, W., "Electrons and Holes in Semiconductors," New York, Prentice-Hall, 1950.

Cross-references: CONDUCTIVITY; ELECTRICAL; CRYSTALLIZATION; ENERGY LEVELS; PHONONS; SOLID-STATE PHYSICS; TUNNELING.

SERVOMECHANISMS

Definition- A servomechanism is a feedback control system in which the difference between a reference input $r(t)$ and some function of a controlled output $c(t)$ is used to supply an actuating signal $e(t)$ to a controller and a controlled system. The actuating signal is amplified in the controller and is used to vary the output of the controlled system in such a manner that the difference between input and output is reduced to zero. A simple block diagram representation of a servomechanism is shown in Fig. 1.

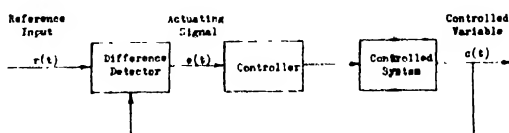


FIG. 1. Block diagram of a servomechanism with unity feedback.

The controlled system may consist of a mechanical structure, a chemical process, a heating system, an electric supply or any system in which a variable can be measured and controlled. The reference input may be a reference level, a sinusoidal or a polynomial function of time, or a discrete, sampled, or programmed set of values. The difference detector is usually matched to the form of the output and input signals. The controller contains signal amplifiers and may also contain power amplifiers which furnish power to an actuating device from an external power source. Actuating devices vary with the controlled system. Actuating devices to position or move mechanical structures may be electromechanical, hydraulic, or pneumatic.

The controller is designed to control a system with known dynamic properties to respond to a specified type of input signal with a specified steady-state and transient performance. The controller may be digital or analog and may be designed for nonlinear or linear operation or may have a linear operating region of actuating signals and a nonlinear saturation zone. Nonlinear systems can be designed with better performance characteristics than linear systems, but linear systems are easier to analyze. A simple linear system with unity feedback will be described.

Linear servomechanisms are analyzed in terms of their transfer function, which can be obtained by taking the Laplace transform of the differential equations of the open loop system with zero initial energy storage. The transfer function is the ratio of the Laplace transform of the output to the Laplace transform of the actuating signal.

$$G(s) = \frac{C(s)}{E(s)}$$

In most cases, $G(s)$ is or can be approximated as a rational algebraic function and can be expressed in the following form.

$$G(s) = \frac{K_n(T_1s + 1)(T_2s + 1) \dots \left[\frac{s^2}{\omega_1^2} + \frac{2\zeta_1}{\omega_1}s + 1 \right]}{s^n(T_as + 1)(T_bs + 1) \dots \left[\frac{s^2}{\omega_a^2} + \frac{2\zeta_a}{\omega_a}s + 1 \right] \left[\frac{s^2}{\omega_b^2} + \frac{2\zeta_b}{\omega_b}s + 1 \right] \dots}$$

The value of the exponent n is used to classify the type of servomechanism. For $n = 0$ (a type 0 servomechanism), the system is often called a regulator or governor, rather than a servomechanism, the distinction being that a servomechanism must contain integration in its transfer function ($n \geq 1$) in order to be able to reduce its steady-state error for a position input to zero.

The control ratio is the closed-loop counterpart of the transfer function.

$$\frac{C(s)}{R(s)} = \frac{G(s)}{1 + G(s)}$$

The ratio of error (or actuating signal) transform to the input signal transform is given by

$$\frac{E(s)}{R(s)} = \frac{1}{1 + G(s)}$$

The characteristic equation for the closed loop system is given by

$$1 + G(s) = 0$$

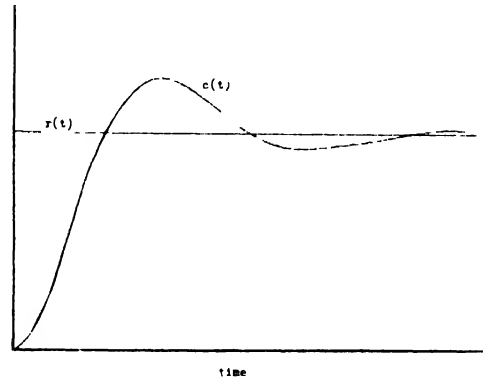


FIG. 2. Transient response of a Type I servomechanism to a step input.

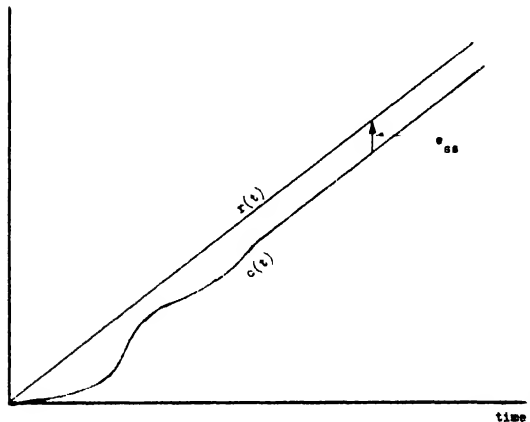


FIG. 3. Response of Type I servomechanism to a ramp input.

For a stable system, there must not be any roots of the characteristic equation in the right half of the complex s plane. A stable system usually has a pair of complex roots much closer to the imaginary axis than any of the other roots. These roots, known as the control poles, are largely responsible for the transient performance of the servomechanism. The controller may contain compensating networks which introduce

zeros to cancel undesirable poles and substitute desirable poles.

Transient performance is often given in terms of the output response to a unit step input—the response frequency, overshoot, and settling time being specified. A value of $\zeta = 0.4$ in the control poles usually limits the transient overshoot to 25 per cent.

Steady-state performance may be specified in terms of the steady-state error during a steady ramp input.

$$r(t) = \omega t$$

For a type I servomechanism ($n = 1$)

$$e_{ss} = \lim_{s \rightarrow 0} sE(s) \\ = \frac{\omega}{K_1}$$

For a type II servomechanism ($n = 2$), the steady-state error for a steady ramp input is theoretically zero.

Servomechanisms can be complicated systems with multiple loops, feedback functions, non-linear or time-varying components, adaptive elements, and undesired random-variable inputs. Special design techniques have been developed to handle these systems.

A very widely used design technique for linear servomechanisms is the Bode plot. The log modulus of the transfer function,

$$\text{Lm}[G(j\omega)] = 20 \log_{10} |G(j\omega)|$$

and the corresponding phase angle of $G(j\omega)$ are

plotted vs frequency on semilog paper. Straight-line asymptotic approximations for the log modulus plots are usually sufficiently accurate. The phase margin, defined as the amount of phase lag less than 180° of $G(j\omega)$ when $\text{Lm}[G(j\omega)] = 0$ decibels, is the basic design quantity. A phase margin of 45° corresponds to control poles with a $\zeta = 0.4$ and a 25 per cent transient overshoot.

A useful tool in the design and analysis of servomechanisms is the root locus plot. The loci of the roots of the closed loop characteristic equation are plotted as the gain is varied. The gain is adjusted until the control poles have a value of ζ corresponding to the desired transient response.

STEPHEN J. O'NEIL

References

- Brown, G. S., and Campbell, D. P., "Principles of Servomechanisms," New York, John Wiley & Sons, 1948.
- Chestnut, H., and Mayer, R. W., "Servomechanisms and Regulating System Design," New York, John Wiley & Sons, 1959.
- James, H. M., Nichols, N. B., and Phillips, R. S., "Theory of Servomechanisms," New York, McGraw-Hill Book Co., 1947.
- Evans, W. R., "Control-System Dynamics," New York, McGraw-Hill Book Co., 1954.
- Truxal, J. G., Ed., "Control Engineers' Handbook," New York, McGraw-Hill Book Co., 1958.
- Truxal, J. G., Ed., "Automatic Feedback Control System Synthesis," New York, McGraw-Hill Book Co., 1955.

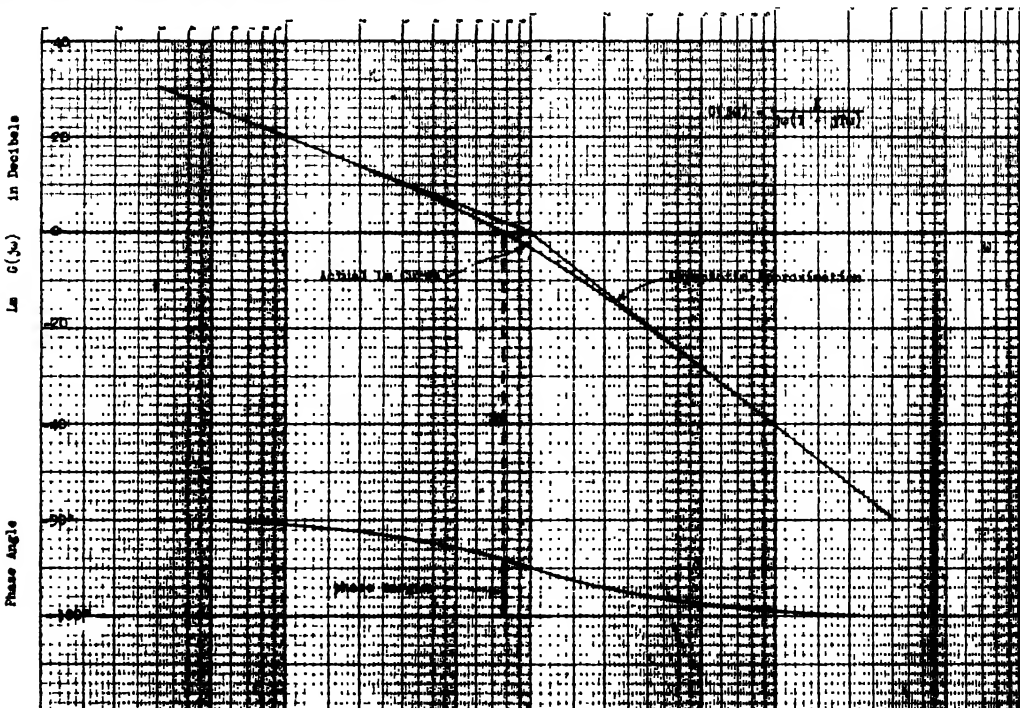


FIG. 4. Bode plots of simple type I servomechanism.

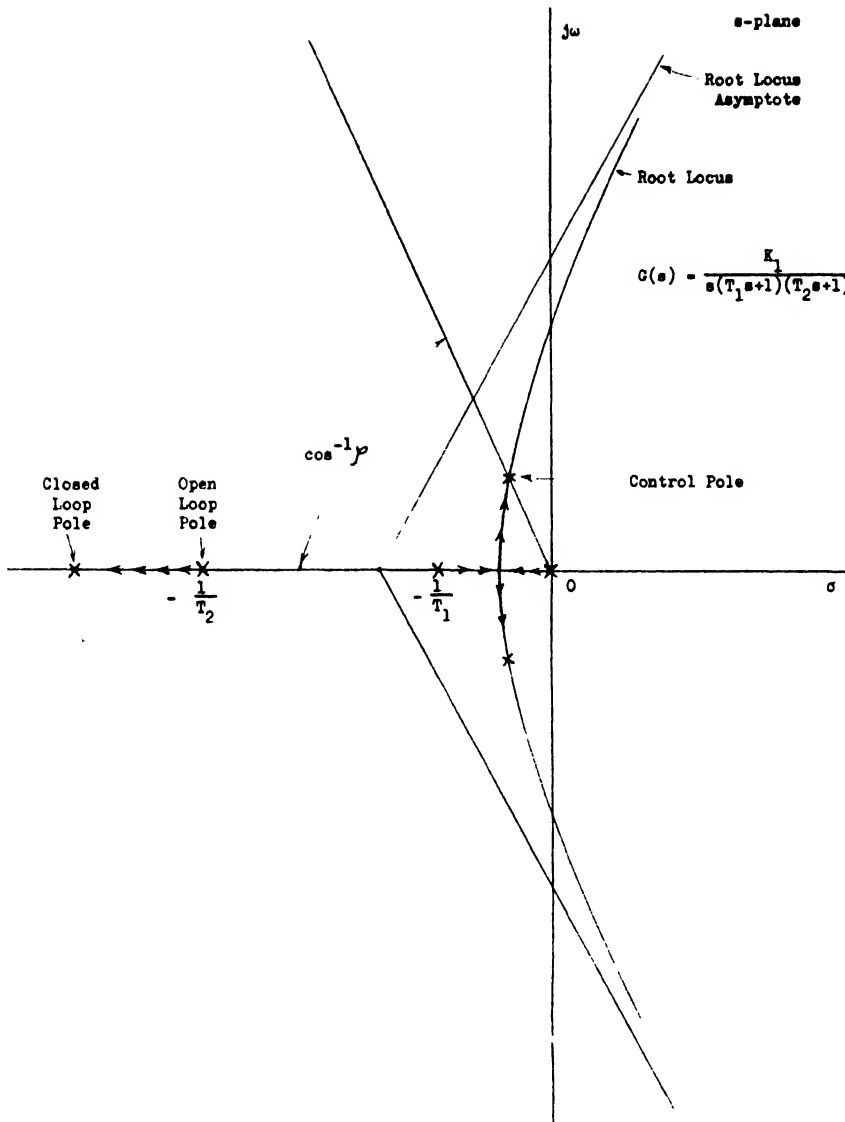


FIG. 5. Root locus plot of typical Type I servomechanism.

- D'Azzo, J. J., and Houpis, C. H., "Feedback Control System Analysis and Synthesis," New York, McGraw-Hill Book Company, 1960.
- Del Toro, V., and Parker, S. R., "Principles of Control Systems Engineering," New York, McGraw-Hill Book Company, 1960.
- Ragazzini, J. R., and Franklin, G. F., "Sampled-Data Control Systems," New York, McGraw-Hill Book Company, Inc., 1958.
- Newton, G. C., Jr., Gould, L. A., and Kaiser, J. F., "Analytical Design of Linear Feedback Controls," New York, John Wiley & Sons, Inc., 1957.

Cross-references: CYBERNETICS, FEEDBACK.

SHOCK WAVES

Infinitesimal disturbances in a fluid medium are propagated with a characteristic speed known as the sound speed. When the restriction on the amplitude of the disturbance is lifted, the linear approximation breaks down and the velocity of propagation becomes dependent on the amplitude of the disturbance. Another feature of this phenomenon is that the forward gradient of the disturbance rapidly steepens until it becomes a discontinuity and propagates as such. A *shock wave* is then a discontinuity in the physical properties of a fluid medium which propagates through

the medium at supersonic velocity without further change. The strength of the shock is defined by the Mach number, the ratio of its velocity to the undisturbed sound speed. Such waves are generated by the detonation of explosive material, by high-speed aircraft and missiles, and by earthquakes and similar natural phenomena.

Since all media are necessarily discrete, a true discontinuity is inconceivable but, as the thickness of the shock transition corresponds to only a few mean free paths in a gas (or internuclear distances in a solid), the transition can be treated as a discontinuity to the same extent that the medium can be regarded as continuous. In comparison with an adiabatic or isentropic change, the shock wave is an irreversible process and hence leads to an increase in the entropy of the material. The pressure, density and temperature of the medium are all raised on passage through the shock and the flow velocity is reduced. The latter is easily understood by observing that, with respect to the moving front, the molecules enter with an ordered flow motion at supersonic speed and the transport processes in the front transform a major fraction of this ordered flow into the random temperature or kinetic motions of the molecules.

The extent to which the various properties change through the transition depends on the magnitude or strength of the shock and on the thermodynamic properties of the fluid. For an essentially incompressible material such as a liquid or solid, the major change normally occurs in the pressure variable, whereas, for a gaseous medium, the most significant change is in the temperature. Although shock waves in solid and liquid materials have been used to study physical properties at high pressures, (see PRESSURE, VERY HIGH), the method is rather limited by the small test times available before the interaction of other wave phenomena which prevent the attainment of thermodynamic equilibrium, and it is in gases that shock waves have proved of most interest.

The detailed behavior of the shock transition is in itself a most important subject for study since shock waves are associated with the flight of supersonic aircraft and with the re-entry of ballistic missiles into the earth's atmosphere. In addition to their own intrinsic interest, shock waves are important for another reason. Since the transition involves the translational motions of the molecules, energy must eventually be transferred into other modes before the system reaches equilibrium. These subsequent RELAXATION processes involve the rotation, vibration, chemical reaction, electronic excitation and even ionization of the molecules if the shock is sufficiently strong. The shock phenomenon thus provides an excellent method for studying energy transfer processes. For chemical reactions in particular, the shock wave provides a source of heat which is essentially instantaneous and is completely homogeneous. Also, provided the thermodynamic properties of the medium are known, the temperature is completely defined by a determination of the shock velocity.

Although shock waves can be created in a large

number of ways, including the detonation of high explosives and the use of wind tunnels, the simplest technique makes use of the *shock tube*, discovered in 1899 by Vieille. A long tube of uniform cross section is divided into two parts by a thin diaphragm and gas is admitted to these at different pressures. If the diaphragm is ruptured in some way, a shock wave is generated in the low pressure gas and a corresponding rarefaction, or expansion, wave in the driver gas. Because the motion is restricted to a single dimension by the containing walls, the strength of the shock does not decrease with distance as it would in a three-dimensional expansion and the relaxation processes become simple functions of distance behind the front. This extremely simple piece of equipment can generate temperatures up to 20 000°K since the strength (or velocity) of the shock depends only on the pressure ratio across the diaphragm immediately prior to rupture and on the thermodynamic properties of the gases in the two sections.

The disadvantage of all shock tube work is that the front moves so rapidly and subsequent wave interactions follow so soon afterwards, that the available testing time is very short, often as low as 100 μ sec. In addition, the total quantity of gas involved is small so that detectors suitable for following the subsequent relaxation processes demand high time resolution and high sensitivity coupled with a suitable physical design which prevents any significant interference with the flow of the gas. Optical devices fit these requirements well, and interferometry, schlieren techniques, ultraviolet, visible and infrared spectrophotometry, and x-ray densitometry have all been used. At the highest temperatures, the shock heated gases become luminous and the shock tube has been used as a spectroscopic source to simulate conditions in stellar atmospheres.

Shock waves can also be created in highly ionized media where the forces are Coulombic in origin and the shocks are termed "collision-less." In this situation, shock waves lie more properly in the realm of plasma physics.

JOHN N. BRADLEY

References

- Bradley, J. N., "Shock Waves in Chemistry and Physics," London and New York, Methuen & Co., Ltd., and John Wiley & Sons, Inc., 1962.
- Gaydon, A. G., and Hurler, I. R., "The Shock Tube in High-Temperature Chemical Physics," London, Chapman and Hall, 1963.
- Courant, R., and Friedrichs, K. O., "Supersonic Flow and Shock Waves," New York, Interscience Publishers, Inc., 1948.
- McChesney, M., "Shock Waves and High Temperatures," *Sci. Am.*, **208**, 109 (1963).

Cross-references: AERODYNAMICS, FLUID DYNAMICS.

SIMPLE MACHINES

Over the years, the ability of man to think has enabled him to find new ways to perform laborious tasks. He developed the lever, ropes, and the inclined plane, modifications of which have resulted in the use of pulleys, wheels and axles, the wedge, and the screw.

Levers. A lever consists of a bar of nearly rigid material, either straight or bent, a fulcrum (F), a weight (W), and a force (P). The components F, W , and P can be applied, in any position relative to each other, to the bar as shown in Fig. 1. Levers are used to move large forces (weights) by means of smaller forces or are used to either amplify or diminish arc motion (Fig. 2). The arms A and B must be perpendicular to the lines of actions (directions) of their respective forces P and W . Then by a balanced moment equation about F,

$$PA = WB$$

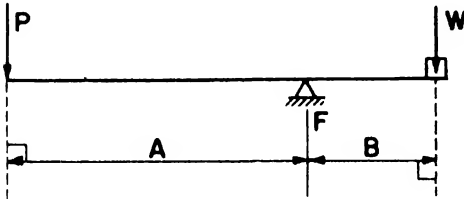


FIG. 1.

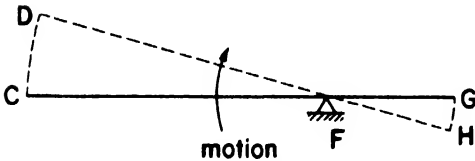


FIG. 2.

The mechanical advantage of a lever, which defines the ability of an available force P to overcome a resisting force W is given by the ratio of W/P or A/B and results in the expression

$$\text{Mechanical advantage} = \frac{W}{P} = \frac{A}{B}$$

Figure 2 illustrates the use of a lever to amplify (or decrease) motion. When bar CG rotates about F, point G moves through the arc GH, and C will move through the arc CD and

$$\frac{CD}{CF} = \frac{GH}{FG}$$

Pulleys. In pulley systems, forces are transmitted by ropes used in conjunction with pulley wheels and axles. As noted from the lever examples, a wheel and axle is an adaptation of a lever rotating about its fulcrum.

Considering a frictionless system of pulleys, the force (pull) in any part of a continuous rope is constant and equal to P . Then, by establishing the number of supporting forces (ropes) and the weight W which is being moved, $nP = W$ where n is the number of supporting ropes. In Fig. 3, two forces (ropes) support the weight W , hence $2P = W$ and $P = W/2$. In Fig. 4, four forces (ropes) support the weight W , then $4P = W$ and $P = W/4$. In Fig. 5, five forces (ropes) support W , then $5P = W$ and $P = W/5$.

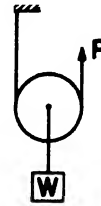


FIG. 3.



FIG. 4.

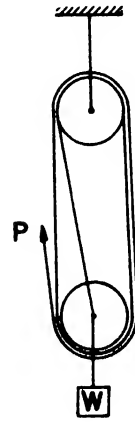


FIG. 5.

The mechanical advantage of a pulley system is the ratio of the weight to be moved to the applied pull in the rope or $W/P = n$, the same as the number of supporting ropes. In Fig. 3, $W/(W/2) = 2$ or for Fig. 4, the mechanical advantage is $W/(W/4) = 4$. In Fig. 5, the mechanical advantage is $W/(W/5) = 5$.

Pulley systems can be analyzed by the use of the work (refer to the section on WORK, POWER AND ENERGY) done by the force P and its relation to the work done by W . The displacement (in Fig. 4) of P is four times that of W . Then the work done by P is $4PS$ where $4S$ is the distance that P moves. The work done by W is WS . The mechanical advantage is $4S/S = 4$ and $4P = W$ as above.

Differential Pulley. The differential pulley makes use of two pulleys of different radii r_1 and r_2 attached to each other and rotating about a common axle. An endless chain connects the dual pulley to a second free pulley wheel as shown in Fig. 6. The chain and the corresponding

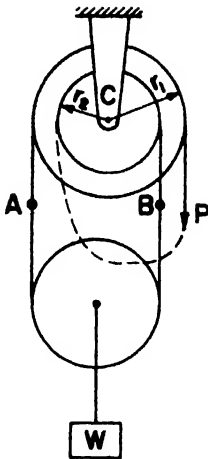


FIG. 6.

teeth on the dual pulleys prevent slipping between the chain and pulleys.

From the previous analysis of pulleys, a force equal to $W/2$ acts in the chain at points A and B. A moment equation about the axle C then gives

$$Pr_1 + \frac{W}{2}r_2 = \frac{W}{2}r_1$$

from which

$$P = \frac{W(r_1 - r_2)}{2r_1}$$

The mechanical advantage of the differential pulley is the ratio W/P .

The Inclined Plane. The inclined plane, as a simple machine, is presumed to be rigid and smooth. The weight (W), which moves along the incline, is a vertically downward force partially supported (N) by the frictionless plane. If the weight W is to be at rest on the incline, Fig. 7, the force system must be balanced in directions normal and parallel to the inclined plane. Balancing the forces parallel to the incline,

$$P = W \frac{h}{L} \quad \text{or} \quad PL = Wh$$

This is equivalent to a work equation where P is displaced a distance L and W is lifted a distance h .

When the force P acts to the left in a horizontal direction, P moves an equivalent distance of b , then

$$Pb = Wh$$

The mechanical advantage, as before, is the ratio $W:P$ or $L:h$ where the force is parallel to the incline. A wedge is equivalent in its analysis to an inclined plane. When t is the thickness of a wedge, $P:W::t:L$.

The Screw. The screw is an inclined plane wrapped around a cylinder in such a way that the height h is parallel to the axis of the cylinder.

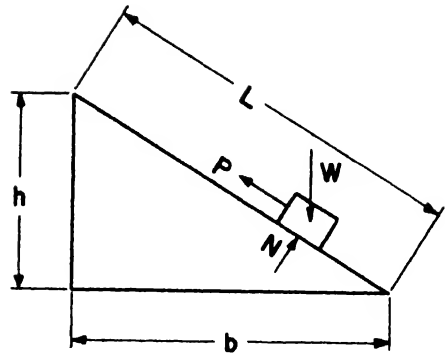


FIG. 7.

If p is the height of travel in one circumference of the screw thread and r is the radius of the thread and friction is neglected, by the work method of analysis

$$P \cdot 2\pi r = Wp$$

The mechanical advantage is

$$\frac{W}{P} = \frac{2\pi r}{p}$$

Gears and Gear Trains. A gear is a wheel with projections uniformly spaced around its circumference. It is usually meshed with a second similar wheel of a different diameter so that the circumferential forces and rotational speeds are different. Figure 8 shows two meshed gear wheels, the teeth of which are not shown. From a balanced moment equation,

$$PR_1 = WR_1$$

then

$$\frac{W}{P} = \frac{R_1}{r_1}$$

and the mechanical advantage is the ratio of W/P or R_1/r_1 .

Gear trains consisting of more than two gear wheels have the same relationships. In Fig. 9, letting the wheels be represented by letters, then

$$PR_1R_2R_3 = WR_1r_2r_3$$

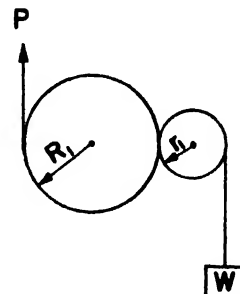


FIG. 8.

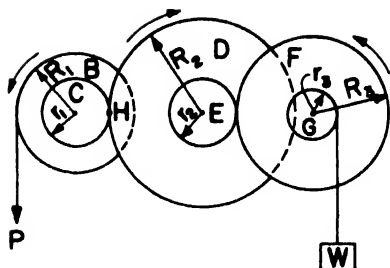


FIG. 9.

and the mechanical advantage is

$$\frac{W}{P} = \frac{R_1 R_2 R_3}{r_1 r_2 r_3}$$

By inspection and understanding that the points of contact, such as point H, have a common speed, the directional relations can be determined. If wheels B and C are rotating counterclockwise, D and E rotate clockwise, and F and G rotate counterclockwise.

Spur gears are those which have their teeth cut parallel to the parallel axes of rotation. Bevel gears are used when their axes of rotation intersect and have their teeth cut on the conical surfaces with their apex at the point of intersection of their axes. Their speed ratios are inversely proportional to their pitch diameters, number of teeth, or their number of revolutions. Referring to Fig. 9,

$$\frac{\text{rpm of } E}{\text{rpm of } F} = \frac{2R_3}{2r_2} = \frac{R_3}{r_2}$$

When a screw meshes with a cogged wheel, it is known as a *worm and worm wheel*. When the worm has a single continuous thread, one revolution of the worm will cause the wheel to rotate equivalent to a circumferential distance equal to the distance between two consecutive teeth. Thus a worm must rotate 48 revolutions if a worm wheel with 48 teeth is to rotate one revolution. The same relationship is true if speeds are considered.

Efficiencies of Simple Machines. The efficiency of a machine is defined as the ratio of the output to the input. The efficiency of a lever is relatively high since the only loss will occur at the fulcrum. Assuming a frictional moment of 5 per cent of the weight moment, then

$$PA = (W + 0.05W)B$$

and the efficiency (per cent) = $(W/1.05W)(100)$. For pulley systems, the efficiencies are rather low because of the frictional resistances at the pulley wheels. A "rule of thumb" states that the weight W shall be increased by 10 per cent for each wheel over which any rope passes. Thus if there are four pulley wheels, the efficiency (per cent) = $[W/(W + 0.4W)](100)$. Differential pulleys have efficiencies approximating 30 per cent due to the large amount of friction present. The efficiency of

an inclined plane is given by the expression

$$\text{Efficiency (per cent)} = \frac{h}{h + fb} 100$$

for the case where the applied pull tends to move the block up the incline (Fig. 7) and f is the coefficient of friction between the incline and the block.

JOHN W. BRENNAN

SKIN EFFECT

For a steady unidirectional current through a homogeneous conductor, the current distribution over the cross section is uniform. However, for an alternating current, the current is not uniformly distributed over the cross section but is displaced more and more to the surface as the frequency increases. For very high frequencies, practically the entire current is concentrated in a thin layer at the surface or "skin" of the conductor. This phenomenon has commonly come to be called the "skin effect."

For a physical explanation of this phenomenon, consider a cylindrical conductor that is carrying a steady unidirectional current. The current distribution over the cross section is uniform. A magnetic field is set up by the current in which the magnetic flux lines within and around the conductor are symmetrical to the conductor axis, i.e., they are in concentric circles. Consider the conductor to be composed of very small circular filaments of equal area which carry equal fractions of the total current. Consider a filament A near the axis of the conductor and another filament B near the surface. The flux linking A is greater than that linking B. If the steady unidirectional current is changed to an alternating current, the magnetic field which is set up by the current must reverse itself periodically with the result that the flux linkages change with time. Self-induced electromotive forces (emf's) are created within the conductor due to these flux changes. These emf's tend to generate currents (eddy currents) which oppose the main current. Because the emf created in filament A is greater than that in B due to a greater flux linkage, the net or resulting current in A becomes less than in B. The current in the conductor is no longer uniform but is displaced toward the surface. The current density becomes greater at the surface and decreases toward the center. This current displacement becomes more pronounced the higher the rate of change of flux.

The total resistance offered by the conductor to alternating current is greater than to a steady unidirectional current because of skin effect. When the current is displaced from the center of the conductor and crowded into the area near the surface, the effective cross section of the conductor is reduced, thereby increasing its resistance. The resistance of the conductor increases with frequency.

The equations which mathematically describe skin effect can be derived from Maxwells' equations for electrodynamic problems. However,

exact solutions have been obtained for only a few simple shapes of conductor. Shapes which have been amenable to analysis include the cylindrical conductor, tubular conductor, flat conductor, surface plated flat, cylindrical and tubular conductor, and coaxial conductor. Even then it is necessary to consider conductors whose material properties do not change with time or temperature.

The first theoretical explanation of skin effect for wires of circular cross section was given by Lord Kelvin in 1887. Early investigators were Kennelly, Laws, and Pierce,² who conducted comprehensive experiments on skin effect, and Dwight³ who obtained solutions for skin effect in tubular and flat conductors. Many other investigations have been made since that time.

The solution to the skin effect equation for a flat or plane conductor carrying a sinusoidal alternating current shows that the current density is maximum at the surface and decreases exponentially in magnitude with distance from the surface into the conductor. Also, as the distance from the surface increases, the current lags in time-phase further and further behind the current at the surface. A quantity $\delta = (2/\sigma\omega\mu)^{1/2}$ is often called the "skin thickness" since it corresponds to the distance from the surface in which the current density drops to $1/e$ of its value at the surface. In this equation, σ is the conductivity, μ is the permeability of the material, ω is the frequency. The resistance and internal reactance of the plane conductor are equal at any frequency. Also, the ac resistance of the conductor is exactly the same as the dc resistance of a plane conductor of thickness δ .

For a cylindrical conductor carrying a sinusoidal alternating current, the solution to the skin effect equation is found in terms of the zero-order Bessel function of the first kind. If the frequency is quite high, the exact formulation in terms of Bessel functions reduces to the simple exponential solution for flat or plane conductors. A useful engineering approximation for the ratio of ac to dc resistance is $R_{ac}/R_{dc} = a/2\delta$ where a is the radius of the conductor and $\delta \leq 0.1a$. The exact ratio in terms of Bessel functions must be used to obtain a solution for low frequencies or small conductors where $0.1a < \delta < a$. At radio frequencies, there is practically no magnetic field inside the conductor, so the reactance is negligible.

The solution to the skin effect equation for tubular conductors is expressed in terms of Bessel functions of the first and second kind. Since the interior portion of a cylindrical conductor carries very little current at high frequencies, thin tubular conductors are often used with a corresponding saving in material. At these frequencies, the tube acts like a solid conductor and effectively serves as an electromagnetic shield.

Conductors are often surface plated for specific applications. Since silver has high conductivity, resonant cavities and wave guides which operate at very high frequencies are silver plated to reduce I^2R loss since most of the current is concentrated

at the surface. Analyses have been made of a number of surface plating combinations to determine what effect such coatings have on the over-all resistance.

The skin effect phenomenon has a very practical application in induction heating. Since the current is concentrated near the surface, a highly selective heating source is created in the surface itself. Advantage is taken of this phenomenon and electromagnetic induction to create a heating method which requires no external heat source and no physical contact with the energy source, the induction coil.

Consider an induction coil carrying alternating current which is wound uniformly around a metal cylinder. Eddy currents are induced in the metal and tend to concentrate near the surface. The problem is similar to that of the cylindrical conductor previously described except that the induced currents are in concentric circles at right angles to the axis of the cylinder instead of being in the axial direction. The solution of the skin effect equation is in terms of the first-order Bessel function instead of the previous zero-order function.

JOHN C. CORBIN, JR.

References

1. Thomson, W., "Mathematical and Physical Papers," Vol. 3, p. 493, Cambridge, Cambridge University Press, 1890.
2. Kennelly, A. E., Laws, F. A., and Pierce, P. H., "Skin Effect in Conductors," *Trans. AIEE*, 34, 1953 (1915).
3. Dwight, H. B., "Skin Effect in Tubular and Flat Conductors," *Trans. AIEE*, 37, 1379 (1918).
4. Ramo, S., and Whinnery, J. R., "Fields and Waves in Modern Radio," Ch. 6, New York, John Wiley & Sons, 1953.
5. McLachlan, N. W., "Bessel Functions for Engineers," Ch. 8, London, Oxford University Press, 1955.
6. Flugge, S., Ed., "Handbuch der Physik," Vol. XVI, "Electric Fields and Waves," p. 182, Berlin, Springer-Verlag, 1958.
7. Moon, P., and Spencer, D. E., "Foundations of Electrodynamics," Ch. 7, New York, D. Van Nostrand Co., 1960.
8. Simpson, P. G., "Induction Heating," New York, McGraw-Hill Book Co., 1960.
9. Corbin, J. C., "The Influence of Hysteresis on Skin Effect," OAR Research Review, Vol. III, 2, April 1964 (U.S. Air Force Publication).

Cross-references: ALTERNATING CURRENTS; CONDUCTIVITY, ELECTRICAL; INDUCTANCE; INDUCED ELECTROMOTIVE FORCE.

SOLAR ENERGY SOURCES

The existence and persistence of biological life on our planet during the last two or three billion years informs us that the solar light intensity has been very steady for at least that length of time.

This in turn points to the fact that the source of solar power must be nuclear energy. Any other source (chemical, gravitational, etc.) would have been exhausted long ago.

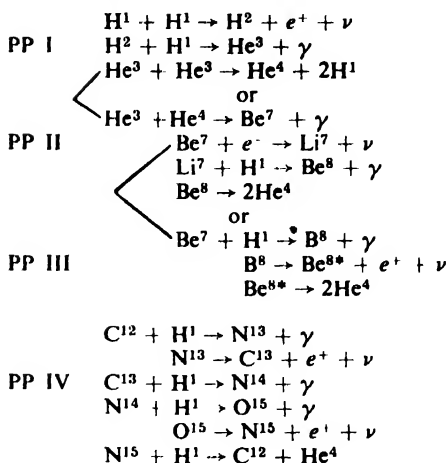
The detailed mechanism responsible for solar energy production was identified some thirty years ago. The energy is liberated when, through various networks of reactions, four hydrogen atoms (H) are combined to form one helium atom (He). The difference of mass between the four hydrogen atoms and the helium atom ($\approx 5 \times 10^{-29}$ gram, corresponding to 26.740 MeV) is then liberated, mostly in the form of gamma rays (about 96 per cent, but also partly in the form of neutrinos (about 4 per cent). The gamma rays (γ) are quickly transformed into heat while the neutrinos (ν) escape immediately and are lost as far as solar heating is concerned.

Some 4.7 billion years ago the mass of gas which was to become the sun was somehow detached from a bigger mass and started to condense under its own weight. From the outside, the stellar mass looked red and was a lot more brilliant than the present sun. The particles composing the gas were mostly hydrogen atoms (about 90 per cent of the atoms), helium (about 9 per cent), then carbon, nitrogen, and oxygen (less than 1 per cent.) and finally traces of many other elements including some metals such as iron. As the contraction proceeded the gas grew hotter. When the temperature reached about one million degrees centigrade, thermonuclear reactions between the atoms in the gas started to occur and to liberate nuclear energy. As the power released by these reactions became comparable to the power radiated away by the sun, the contraction stopped, the temperature remained constant and nuclear energy took over the burden of keeping the sun warm. The period of contraction had lasted about 10^7 years. By then the sun had taken its present yellowish appearance and had approximately its present-day brilliance. It had become a so-called main-sequence star.

The transformation of hydrogen into helium then mainly involved the following set of reactions. First in a collision between two hydrogen atoms (H^1) a nuclear reaction takes place and an atom of deuterium (H^2 , heavy hydrogen) is formed. This reaction is by far the slowest that we shall meet. It essentially governs the rate of solar energy generation. Next the deuterium reacts with another hydrogen to form an atom of helium 3 (He^3 , the light isotope of helium); then two helium 3 thus formed react together to form one helium 4 (He^4) isotope and to release two hydrogen atoms. (The collision between helium 3 and hydrogen yields nothing.) The chain (called the proton-proton or PP I chain) is summarized in Table I.

As the concentration of helium increases, another reaction becomes important, the reaction between one helium 3 and one helium 4, resulting in the production of one beryllium 7 (Be^7) isotope. In this reaction, the beryllium atom captures a free electron in the stellar gas and decays into a lithium 7 (Li^7) atom. Then the

TABLE I. THE PROTON-PROTON CHAINS



γ represents gamma rays; ν represents neutrinos

lithium 7 atom captures a hydrogen atom, thereby forming a beryllium 8 (Be^8) atom. This atom is highly unstable. It rapidly breaks into two helium 4. This forms the PP II chain (see Table I).

In the present sun, 40 per cent of the energy comes from the PP I chain and 56 per cent from PP II. The central temperature is about sixteen million degrees and the central density 180 g/cc. (The density of water is 1 g/cc.)

A third chain is started if the beryllium 7 atom absorbs a hydrogen atom before it has time to decay. A boron 8 (B^8) atom is thus formed which quickly decays to beryllium 8. The beryllium 8 breaks apart releasing two helium 4 atoms. The contribution of this chain to the total energy generation is negligible (0.05 per cent). Its interest lies in the fact that with the decay of the beryllium 8, a neutrino (ν) with a mean energy of about 7 MeV is emitted. The other branches are also accompanied with neutrino emission but with mean energy less than 2 MeV. High-energy neutrinos are far easier to detect than low-energy neutrinos. Later we shall discuss a project already underway which should bring about the detection of solar neutrinos.

Finally, the fourth mode of hydrogen to helium conversion in the sun is the famous carbon cycle. First an atom of carbon 12 (C^{12}) captures hydrogen to produce nitrogen 13. Then nitrogen 13 (N^{13}) decays to carbon 13. Carbon 13 (C^{13}) forms nitrogen 14 (N^{14}) and oxygen 15 (O^{15}) by absorbing successively two hydrogens. Oxygen 15 decays to nitrogen 15 (N^{15}). Nitrogen 15 captures a last hydrogen and breaks into carbon 12 and helium 4. Carbon 12 is thereby returned to the gas, ready to start the cycle again.

Until a few years ago, this cycle was thought to be the main source of energy generation in the sun. Better solar models made with more accurate evaluation of the opacity of the solar material have shown that only about 4 per cent of the sun's brightness comes from this cycle.

As mentioned before, the sun has been in the process of converting hydrogen into helium for the past 4.5 billion years or so. Now the number of helium atoms in the center reaches 30 per cent from the 10 per cent that was there originally. As this number keeps on increasing, a bigger and bigger fraction of the solar energy will come from the carbon cycle. In a few billion years, as the core gets depleted of hydrogen, the sun will slowly start warming up again. A thin shell of hydrogen surrounding the helium core will get hot enough to generate energy, this time almost entirely through the carbon cycle. As this process goes on, the envelopes will start expanding at a rather fast rate; the sun will be entering its red giant phase. The surface, somewhat cooler than it is now, but still at temperature above the melting and vaporizing point of all material, will gradually swallow Mercury, Venus, the earth, and possibly Mars and Jupiter. Humanity will have learned to escape the doomed solar system or it will be destroyed. However since these events will take place in a few billion years, there is still time . . .

Eventually the central temperature will reach the one hundred million degree mark and the central helium will be ignited. The collision of two helium 4 nuclei induces the formation of one beryllium 8 atom which, however, soon breaks apart into its initial components.

However, as this sequence of combination and dissociation goes on (similar to processes in atomic and molecular gases), an equilibrium concentration of beryllium builds up. Occasionally one of these atoms will hit another helium 4 and form a carbon atom. The over-all reaction is labelled $3\text{He}^4 \rightarrow \text{C}^{12}$. It will provide energy for the sun for a few tens of millions of years.

Eventually, when helium is exhausted the sun will keep on warming its center and will burn its carbon and then its oxygen (both of these elements are products of the helium burning stage). Soon it will have no more energy sources to draw on. It will start cooling and contracting—first becoming a white dwarf, then shining no more, and losing all contact with the external world except through its gravitational field.

How and how well do we know that? What is the basis of our exploration of the past, the present, and the future of our sun? In our hands we have two different tools: physics and astronomy. Laboratory experiments are the foundation of our understanding of matter at our scale—the atomic scale and the nuclear scale. From these experiments, theories have been constructed which we believe can be applied in extraterrestrial settings. Astronomical observations are the basis of our cognizance of extraterrestrial settings.

From observations made on the sun we can obtain its mass, its radius, its luminosity, and the chemical composition of the surface. From other individual stars, we get even less information. However from statistical studies made on stars, and in particular on star clusters, we obtain most of our knowledge of stellar evolution (the same way one can investigate the pattern of growth of trees by walking in a forest).

The two main observational characteristics of a star are its color and its absolute brightness (i.e., how much energy it pours out). Families of stars of similar or related brightness and color have been identified for a long time. The most important groups of stars are called main sequence stars, red giants, and white dwarfs. After many years of patient labor involving the working out of stellar models from our knowledge of hydrodynamics, atomic and nuclear physics, and the comparison of these models with actual stars or groups of stars, the pattern of evolution has become apparent. After a period of contraction, a star spends most of its life as a main-sequence star (such as the present sun), then it becomes a red giant, and eventually after shedding out some of its mass, it becomes a white dwarf.

The story of our sun described in the first pages of this text is based on such evidences. How solid such evidences are, one never really is certain, but in view of the enormous amount of observations properly explained this way, one feels confident.

Another piece of information about the sun should become available within a few years. As mentioned before, a fraction of the solar energy is used to generate neutrinos. These neutrinos come from the hot furnace in the center of the sun and escape immediately. A program for the detection of these neutrinos has been underway for some time. The neutrinos are extremely difficult to detect and consequently the experiment is a highly involved one. If the experiment is successful, it will give us first hand information about the solar interior. First we shall get direct confirmation of the nuclear origin of solar energy; second we shall obtain a very accurate determination of the solar temperature. Should this measurement be very different from our present estimation we would have to worry about our knowledge of stellar structure and possibly be on the way toward exciting new discoveries.

HUBERT REEVES

References

- Schwarzschild, M., "Structure and Evolution of Stars," Princeton, N. J., Princeton University Press, 1958.
- Stromgren, B., "The Sun as a Star," in "The Sun," Chicago, The University of Chicago Press, 1953.
- Rudaux, L., and DeVancouleurs, G., "Larousse Encyclopedia of Astronomy," New York, Prometheus Press, 1962.
- Aller, L. H., "Astrophysics," Vol. II, New York, The Ronald Press Co., 1954.
- Reeves, H., "Stellar Energy Sources," in Aller, L. H., and McLaughlin, D., Eds., "Stars and Stellar Systems," Chicago, University of Chicago Press, 1964.
- Schatzman, Evry., "Origine et Evolution des Mondes," Paris, Editions Albin Michel, 1960.
- Menzel, D. H., "Our Sun," Philadelphia, Blakiston Co., 1958.

Cross-references: ASTROPHYSICS, NEUTRINO, NUCLEAR REACTIONS, NUCLEONICS, SOLAR PHYSICS.

SOLAR ENERGY UTILIZATION

The use of solar energy is likely to become of increasing importance both on earth, because of the rapidly diminishing supplies of fossil fuels, and in man's expanding adventures into space. The problem of economically using solar energy for various terrestrial needs has challenged us for many years, but there are still only a few economically attractive applications. Technological advances may well change this situation and help satisfy mankind's growing appetite for more and more energy. The launching of the first earth satellite opened up a completely new era for solar energy utilization. For space applications, reliability and weight are generally much more important than the cost of the power system itself. Silicon photovoltaic power supplies already have found extensive use on satellites and space vehicles.

The Sun as an Energy Source. Our sun is a typical main sequence dwarf of spectral class G-2. In almost every respect (size, luminosity, mass, etc.), it is an average star. The mean distance from the earth is 93 004 000 miles, at which distance the sun subtends an angle of 31 minutes 59.3 seconds.

The sun's total radiation output is approximately equivalent to that of a blackbody at 10 350 R (5750 K). However, its maximum intensity occurs at a wavelength which corresponds to a temperature of 11 070 R (6150 K) as given by Wien's displacement law. Figure 1(a) presents the intensity of solar radiation outside the earth's atmosphere as a function of wavelength. The total irradiance at the mean sun-earth distance is approximately 442 Btu/hr ft² (0.140 watts/cm²). Figure 1(b) shows how the energy in the sun's spectrum is distributed.

Sunlight passing through the atmosphere suffers both absorption and scattering. The solar irradiance reaching the earth depends upon the total air mass through which the sun passes and, of course, weather conditions. Figure 1(c) shows a typical plot of the solar irradiance at the earth's

surface. For some applications, the radiation from the sky is also important. The day sky exhibits wide variations in brightness and in spectral distribution, depending upon the sun's position, weather conditions, and the "receiver's" orientation.¹ Table 1 presents typical values of direct and diffuse solar radiation for several different atmospheric conditions. A great deal of study has been made of the variation of the solar irradiance at the earth's surface with weather, atmospheric conditions, zenith angle of the sun, etc.^{2,3}

Uses for Solar Energy. Table 2 lists most types of application of solar energy of interest today. They are grouped according to the form of energy required by the application. As an example, supplying electrical power for space vehicles requires a conversion from solar to electrical energy which may be a one-step process with photovoltaic cells (solar cells) or a multistep process. A typical multistep process might involve conversion of solar energy to thermal energy by using a parabolic mirror to heat a working fluid in a boiler, then conversion to mechanical energy using Stirling engine, and then conversion to electrical power using a generator. Reference 4 discusses each of these applications and comments on their economic feasibility. The interested reader is directed to this and other references listed at the end of this article for these details.

Considerations in the Conversion of Solar Energy to Other Forms. Table 2 indicates that most of the applications of solar energy require that the energy be delivered in one of four forms: thermal, electrical, mechanical, or chemical. It is of interest, then, to examine the techniques and limitations of converting solar energy to these other forms.

Table 3(a) shows several methods of directly converting solar energy to these other forms. Table 3(b) shows other conversion steps that might be used in a multistep process. The devices described in Table 3(b) make use of processes which are well known and not restricted to solar energy utilization. Those in Table 3(a), on the

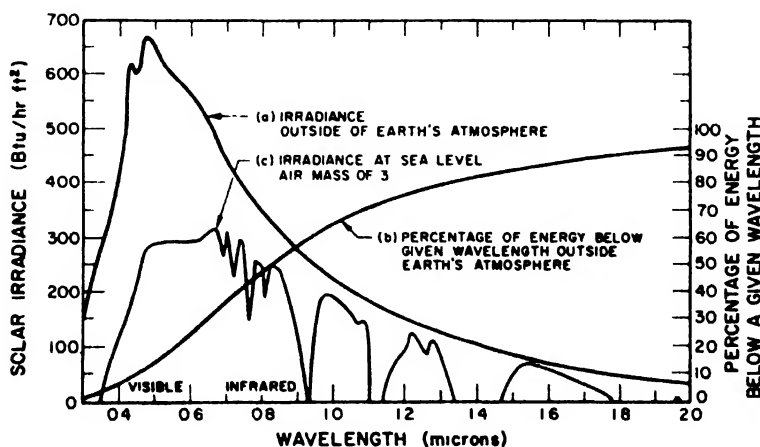


FIG. 1. Spectral distribution of solar energy.

TABLE 1. VALUES OF DIRECT AND DIFFUSE SOLAR RADIATION*

Solar Altitude α (degrees)	Optical air-mass path, \dagger m $\sim \csc \alpha$	Standard, Cloudless Atmosphere			Industrial, Cloudless Atmosphere			Through Complete Overcasts, Blue Hill, Average Total Insolation on Horizon			
		Direct, Perpen- dicular Radiation I (B/hr ft ²)	Diffuse on Horizontal I_z Difference (B/hr ft ²)	Total on Hori- zontal $I_z + I$ (B/hr ft ²)	Direct, Perpen- dicular Radiation I (B/hr ft ²)	Diffuse on Horizontal I_z Dif- ference (B/hr ft ²)	Total on Hori- zontal $I_z + I$ (B/hr ft ²)	Cirro- stratus W_z (B/hr ft ²)	Alto- cumulus W_z (B/hr ft ²)	Strato- cumulus W_z (B/hr ft ²)	Fog W_z (B/hr ft ²)
5	10.39	67	7	13	34	9	12	—	—	—	—
10	5.60	123	14	35	58	18	28	—	—	15	10
15	3.82	166	19	62	80	24	45	50	35	25	15
20	2.90	197	23	90	103	31	64	70	50	35	20
25	2.36	218	26	118	121	38	89	95	65	40	20
30	2.00	235	28	146	136	44	112	120	75	50	25
35	1.74	248	30	172	148	48	133	145	90	60	30
40	1.55	258	31	197	158	52	154	165	105	70	35
45	1.41	266	32	220	165	55	172	185	115	80	40
50	1.30	273	33	242	172	58	190	205	130	85	40
60	1.15	283	34	279	181	63	220	235	150	100	45
70	1.06	289	35	307	188	69	246	260	160	110	50
80	1.02	292	(35)	(322)	195	—	—	—	—	—	—
90	1.00	294	(36)	(328)	200 \ddagger	—	—	—	—	—	—

* Data assembled by F. A. Brooks. "B" in the headings stands for British thermal units (Btu).
† Smithsonian Meteorological Tables, 6th rev. ed., p. 422, 1951.
‡ 192 would be more consistent with the curve from 70' down.

TABLE 2. SOME APPLICATIONS OF SOLAR ENERGY

Form of Energy Required	Application
Thermal	Concentration of brine
	Cooking
	Cooling and refrigeration
	Dehumidifying of buildings
	Distillation of water
	Drying of materials, fruits, grain, etc.
	Heating of buildings
	Heating of water
Electrical	Heating of materials to high tem- peratures
	Salt making
	Supplying power for special uses
Mechanical	Supplying space vehicle power
	Pumping water
Chemical	Producing food
	Producing fuel
Optical	Natural lighting
	Phosphorescent markers

lenses to achieve high flux densities (concentrating collectors) while others do not (flat-plate collectors).

Flat-plate Collectors. The most common flat-plate collector consists of a metal plate painted black on the side facing the sun and thermally insulated on the edges and back. Above the plate, spaced an inch or so apart, are one or more glass or plastic covers to reduce upward heat losses. The absorbed energy is transferred to water or some other working fluid in tubes which are in thermal contact with the absorber plate or by circulating air past the absorber. Figure 2 displays several flat-plate collector designs.⁵ Design B is the conventional type described above; A uses air instead of water for heat transfer. In C, the glass-shingle collector, heat is absorbed on the blackened bottom third of tilted glass plates

TABLE 3. MATRIX OF A FEW ENERGY
CONVERSION DEVICES

(a) Direct Conversion from Solar Energy	
To	From Solar
Chemical	(Photochemical reactions)
Electrical	Photovoltaic cell
Mechanical	Photon "sail"
Thermal	Flat-plate collector Concentrator-boiler

other hand, are of prime interest to the present discussion.

Conversion of Solar Energy to Thermal Energy. Many of the systems which utilize solar energy first collect the energy and its heat. A solar heat collector intercepts radiation, converting this to thermal energy, and transfers this heat to a working fluid. Some collectors use mirrors or

(b) Other Conversion Steps

From \ To	Chemical	Electrical	Mechanical	Thermal
Chemical	---	Electrolysis	(Mechanical activation of chemical processes)	(Endothermic reactions) (Thermal dissociation)
Electrical	Fuel cell battery	---	Generator	Thermopiles Thermionic diodes MHD devices
Mechanical	(Equilibrium volume and pressure)	Motor Solenoid	---	Turbine Positive displacement engine
Thermal	(Combustion)	Resistance heaters	Friction brake	---

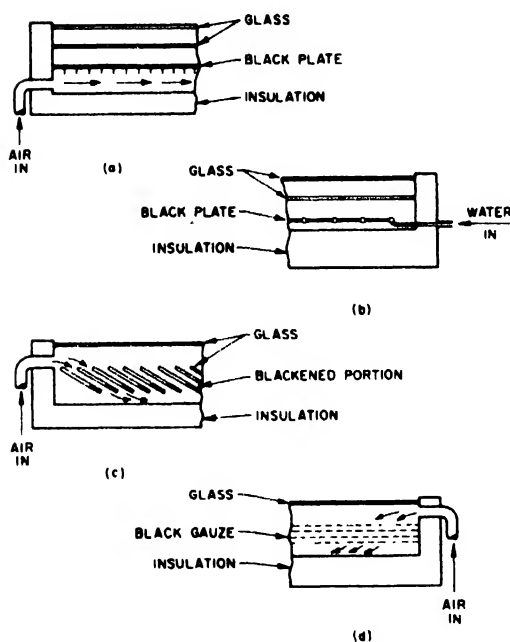


FIG. 2. Four flat-plate collector designs.

and transferred to the air drawn down through these shingles. In D, solar energy is absorbed on black gauze. The flat-plate collector is a simple, rugged device which, without orientation mechanisms, can efficiently collect solar energy at moderate temperature levels.

The results of some comparative calculations of conventional collectors are shown in Fig. 3. The emissivity and absorptance of the absorber plate was assumed to be 0.96, and the curves were plotted for zero, one, two, and three cover plates of typical single-strength window glass. It is seen that for high efficiencies (with correspondingly low collector temperatures), it might be preferable to have only one or two covers. As an example, a two-cover collector is preferable to a three-cover collector out to a temperature of about 200°F (for the case analyzed);

it is only at higher temperatures that the decrease in upward losses made by using one more cover is greater than the transmission loss of that cover.

Concentrating Collectors. An optical system using lenses or mirrors can be used to concentrate solar radiation into a very small area. If this energy is received into a cavity or absorbed on a metal plate, heat is generated and very high temperatures may be obtained.

The concentration ratio C is defined as the ratio of the flux density within the image of the sun formed by the optical system to the actual flux density reflected from the mirror. For a parabolic mirror, C is given by⁷

$$C = 46.1 \times 10^3 \sin^2 \theta \quad (1)$$

where θ is the rim angle of the mirror as defined by the sketch in Fig. 4. Figure 4 plots Eq. (1) and also a parameter called concentration efficiency which is defined as the ratio of the power received within the sun's image to the total power reflected by the mirror. For the parabolic mirror, the concentration efficiency is given by

$$\eta_c = \left(\frac{1 + \cos \theta}{2} \right)^2 \quad (2)$$

The concentration ratio and the concentration efficiency are strongly influenced by the geometric perfection of the mirror and the location of the absorber with respect to the focus of the mirror.⁸

Consider an example where the energy reflected from a parabolic mirror is collected by an ideal cavity receiver having an opening equal to the image diameter of the sun. The rate at which heat energy can be withdrawn from the cavity, P , can be calculated by subtracting the radiative heat losses from the input flux. The result is:

$$P = \eta_c A \left[rH - \frac{\sigma T^4}{c} \right] \quad (3)$$

where A is the projected area of the mirror, r is its reflectance and H is the solar irradiance. T is the temperature of the cavity. Figure 5 uses Eq.(3) to plot the ratio of P/A as a function of cavity temperature for several rim angles. The

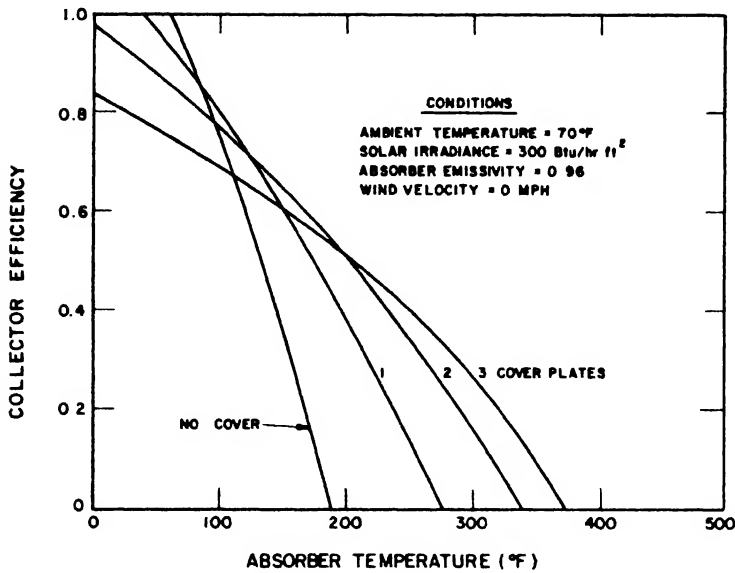


FIG. 3. Calculated performance of flat-plate collectors showing effect of multiple glass covers.

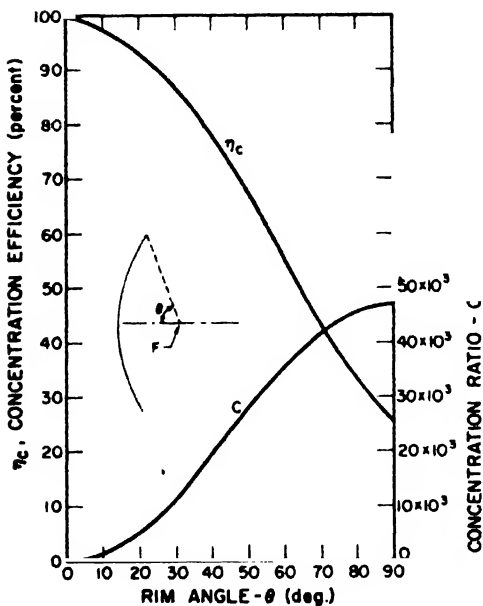


FIG. 4. Concentration ratio and concentration efficiency as a function of rim angle for a paraboloidal reflector.

intersection of the curves with the abscissa show the maximum attainable temperature for various rim angles, corresponding to $P = 0$ in Eq. (3).

A variety of optical systems have been studied for use as solar concentrators with various absorbers.⁹⁻¹¹ The parabolic mirror-cavity receiver example, however, is indicative of the ultimate performance attainable.

Conversion of Solar Energy to Electrical Energy. The PHOTOVOLTAIC EFFECT, particularly in silicon, has become very important to solar energy utilization. In the silicon "solar cell," photons create hole-electron pairs by removing electrons from valence bonds. This occurs in the junction region between *p*-type and *n*-type silicon where the electric field causes a current to flow in the cell. Equilibrium is restored by charge flowing around an external circuit through a load resistor.¹²

Each silicon nucleus shares its four valence electrons with neighboring nuclei forming a stable tetrahedral crystal. In the junction region, most of the electrons are in the lower filled band, below the forbidden energy band. The electron may be thought of as either bound to a crystal lattice region or, with the addition of enough energy, free to move about and conduct electricity. The width of the forbidden region, or energy gap, represents the threshold energy necessary to remove an electron from the bound position to the conduction band. For silicon, it is approximately 1.1 eV. It follows that only photons of 1.1 eV or over can create hole-electron pairs in silicon. Higher-energy photons can create hole-electron pairs with the excess energy being dissipated as heat (see ENERGY LEVELS and SEMI-CONDUCTORS).

Since one hole-electron pair is created for each photon possessing sufficient energy to do so, the ultimate efficiency that might be achieved depends upon the threshold energy at which the hole-electron conversion takes place and upon the spectrum of the radiation. Figure 6 is the plot of the maximum power conversion efficiency as a function of threshold level. It assumes that all of the photons whose energy is greater than or equal to the threshold energy are converted into

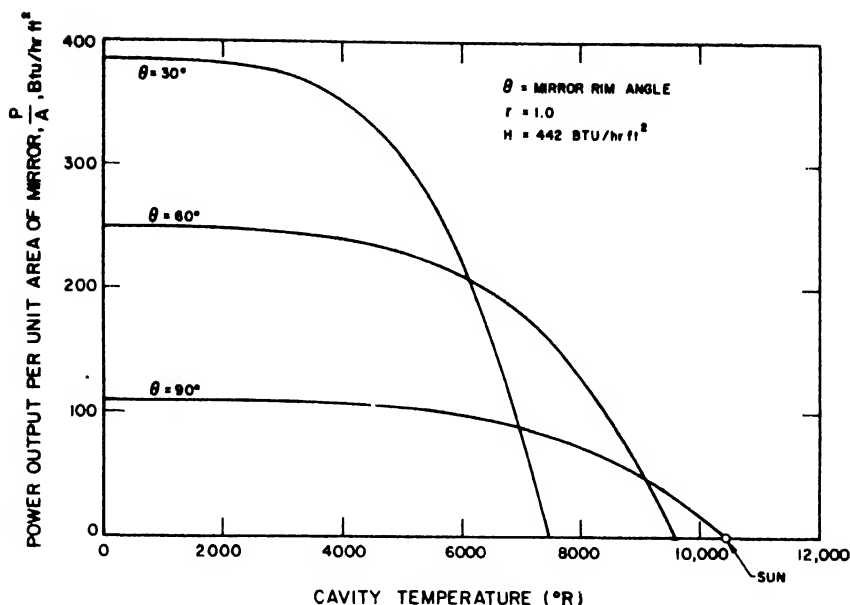


Fig. 5. Maximum power output from an ideal parabolic mirror-cavity absorber as a function of operating temperature.

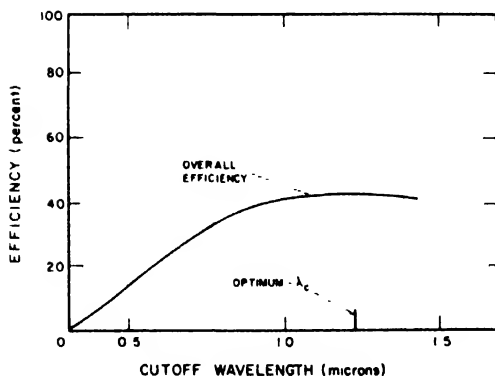


Fig. 6. Maximum solar conversion efficiency of an ideal quantum converter.

hole-electron pairs at the potential of the energy gap.*

Some of the factors which limit the efficiency in a practical cell include:

*Several schemes have been proposed to get around the band gap limitation by using three or more semiconductors sandwiched with the widest-gap semiconductor on top—utilizing the highest-energy photons, a medium-gap material using the next lower energy, etc. It has also been proposed that it might be possible to continuously vary the material so that a variable band gap is achieved. Other schemes which utilize dichroic mirrors to separate the spectrum into two beams of different spectral content to shine on two separate sets of cells with different band gaps have also been proposed.

- (1) Some hole-electron pairs recombine before they can be separated by the field in the junction,
- (2) Reflection losses from the front surface,
- (3) Loss due to electrical resistance within the cell,

- (4) Loss due to leakage across the barrier (diode current).

In full sunlight a good silicon cell will develop approximately 0.6 volt, open circuit, and will operate at a conversion efficiency of 10 to 12 per cent.

Figure 7 is a photograph of one of the solar panels used on the Ranger Spacecraft developed for NASA's Jet Propulsion Laboratory. The panel contains 72 strings of 68 cells each to provide 100 watts output at approximately 30 volts. The use of optical filters to cover the cells results in a lower cell temperature, hence higher efficiency, and provides improved resistance to radiation damage.¹³ It is also possible to improve the power output per unit weight of a solar panel by using lightweight mirrors to concentrate sunlight on the cells.¹⁴

Conversion of Solar to Mechanical Energy. Solar radiation incident upon a reflecting surface exerts a radiation pressure of approximately 0.8×10^{-4} dynes/cm² or 1.7×10^{-7} lb/ft². Although impractical for terrestrial applications, the use of radiation pressure for space propulsion (solar sailing) may prove to be worthwhile for some missions.¹⁵

Conversion of Solar Energy to Chemical Energy. It has been estimated that only a few tenths of a per cent of a year's supply of incident solar energy is stored in an average farm crop. Research into the mass production of algae has

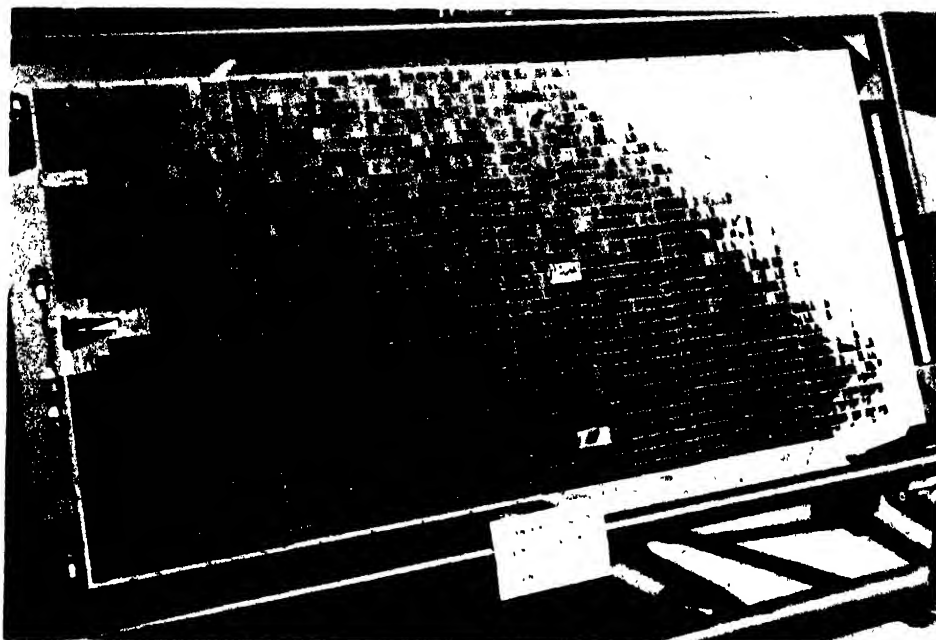


FIG. 7. Ranger spacecraft solar cell panel. (Photograph courtesy Electro-Optical Systems, Inc.)

been directed toward tenfold greater yields; research in other types of plants, toward the end of increasing photosynthetic efficiency, is also being conducted.¹⁶

The existence of photosynthesis raises the hope that other photochemical reactions will be found which can be effectively used to obtain chemical energy from the sun. What is needed is a reversible endothermic photochemical reaction which can utilize a large portion of the sun's spectrum. Most photochemical reactions evolve heat instead of absorbing it, or they reverse so quickly that the energy is lost even during exposure to sunlight.

Water is photochemically decomposed into hydrogen and oxygen by ultraviolet light, though at very low efficiency.¹⁷ A fuel cell could be used to obtain electricity from the stored gases, or mechanical energy could be obtained in an engine. While this process is not presently of economic importance, it serves to demonstrate the possibilities. Unfortunately, 50 per cent of the sun's energy is at wavelengths too long to be of much use for photochemical reactions, since the energy per photon is too low. All molecules which undergo photochemical change have a minimum threshold energy required to create a bond rupture. If a quantum is absorbed with energy above this threshold, the excess energy is dissipated (as kinetic, vibrational, etc.) and usually does not contribute to the conversion to chemical energy. The maximum theoretical efficiency of conversion of solar to chemical energy can be evaluated with the help of Fig. 6 providing the cutoff wavelength is known.

At the present time, there is not one system known which comes close to fulfilling all of the requirements for the ideal energy converter; however, important advances are being made and some chemical systems show a great deal of promise.¹⁸

A. M. ZAREM
DUANE D. ERWAY

References

1. Kruse, P. K., McGlauchlin, L. D., and McQuistan, R. B., "Elements of Infrared Technology," pp. 76, 77, New York, John Wiley & Sons, 1962.
2. Brooks, F. A., and Miller, W., "Availability of Solar Energy," in Zarem, A. M., and Erway, D. D., Eds., "Introduction to the Utilization of Solar Energy," pp. 30-57, New York, McGraw-Hill Book Co., 1963.
3. Landsberg, H. E., "Solar Radiation at the Earth's Surface," *Solar Energy J.*, 5(3), 95 (1961).
4. Zarem, A. M., and Erway, D. D., "Introduction to the Utilization of Solar Energy," New York, McGraw-Hill Book Co., 1963.
5. Nollel, H. C., "Residential Uses of Solar Energy," *Proc. World. Symp. Appl. Solar Energy, Phoenix, Ariz.*, 1955, 103 (1956).
6. Erway, D. D., "Collection of Solar Energy," in Zarem, A. M., and Erway, D. D., Eds., "Introduction to the Utilization of Solar Energy," pp. 89-100, New York, McGraw-Hill Book Co., 1963.
7. Hiester, N. K., Tietz, T. E., Loh, E., and Duwez, P., "Theoretical Considerations on Performance of Solar Furnaces," *Jet Propulsion*, 27, 507 (1957).

8. Duwez, Pol, "Concentration of Solar Energy," in Zarem, A. M., and Erway, D. D., Eds., "Introduction to the Utilization of Solar Energy," pp. 121-123, New York, McGraw-Hill Book Co., 1963.
9. Fisher, J. H. *et al.*, "An Analysis of Solar Energy Utilization," Vol. II, pp. 45-98, Wright Air Development Center Report 59-17, 1959 [available through the Defense Document Center (DDC)].
10. McClelland, D. H., "Solar Concentrators for High Temperature Space Power Systems," in Snider, N. W., Ed., "Space Power Systems," p. 129, New York, Academic Press, 1961.
11. Giutronich, J. E., "The Design of Solar Concentrators using Toroidal, Spherical, or Flat Components," *Solar Energy J.*, 7(4), 162-166 (1963).
12. Prince, M. B., "Silicon Solar Energy Converters," *J. Appl. Phys.*, 26, 534-540 (1955).
13. Hamilton, Robert C., "Ranger Spacecraft Power System," in Snider, N. W., Ed., "Space Power Systems," p. 19, New York, Academic Press, 1961.
14. Menetrey, W. R., "Space Applications of Solar Energy," in Zarem, A. M., and Erway, D. D., Eds., "Introduction to the Utilization of Solar Energy," pp. 373-378, New York, McGraw-Hill Book Co., 1963.
15. Garwin, Richard L., "Solar Sailing—A Practical Method of Propulsion within the Solar System," *Jet Propulsion*, 28(1), 188 (1958).
16. Survey for publication in *Sun at Work*.
17. Daniels, F., and Duffie, J. A., "Solar Energy Research," pp. 119-219, Madison, Wisconsin, University of Wisconsin Press.
18. Rabinowitch, E., "Photochemical Utilization of Light Energy," *Solar Energy J.*, 52 (September 1961, special issue).

Cross-references: ENERGY LEVELS; INFRARED RADIATION; PHOTOCHEMISTRY; PHOTOVOLTAIC EFFECT; REFLECTION; SEMICONDUCTORS; SOLAR ENERGY SOURCES; SOLAR PHYSICS, WORK, POWER, AND ENERGY.

SOLAR PHYSICS

The main activity of solar physics is the interpretation of the observed flow of energy away from the sun. Most of the sun's energy output appears as a nearly constant flux of electromagnetic radiation in the photographic, visual, and infrared regions of the spectrum, but extraordinary variations are characteristic of the x-, ultraviolet, and radio radiation. A small part of the sun's energy flows outward in highly variable streams of particles (mainly protons and electrons). Since the sun appears as a rather large disk, the observations forming the bases for the solar physicist's interpretations often refer to small areas of the solar surface, as well as the integrated radiation and particle streams from the entire sun. A reasonably good solar telescope can subdivide the apparent disk of the sun into about three million smaller elements of area. Streams of particles and radiation from each of these elements should be recorded, with all possible detail and precision, nearly continuously for a complete observational record. A close approach to such

observational perfection probably is unattainable, perhaps it is unnecessary, but the fragmentary and incomplete nature of all solar observations continues to be a serious barrier in the search for a satisfactory general theory of the sun. Some progress has been made, but until an adequate theory is developed, it is best to consider the physics of the sun by reviewing the observations of possible importance.

The radius, mass, and luminosity of the sun are fundamental in the physical interpretation of the sun. Luckily, they can all be deduced with reasonable directness from observations. In principle, the radius of the sun can be obtained directly from angular and linear observations made on earth, but a number of practical difficulties limit the attainable precision. Somewhat devious methods, invoking gravitational theory, give more consistent results that converge on the value,

Radius of the sun, $R = 6.9598 \pm 0.007 \times 10^{10}$ cm

The mass of the sun is also derived from the application of gravitational theory: first, to the measurement of the mass of the earth; then, by way of the moon and planets, to the sun; with the result,

Mass of the sun, $M = 1.989 \pm 0.002 \times 10^{33}$ gram

A measurement of the total radiation received at the earth's distance from the sun is the basic observation from which its luminosity can be found (once the distance to the sun, and the radius of the sun are known). This measurement is extremely difficult and is seriously distorted by the earth's atmosphere, but the value

Luminosity of the sun, $L = 3.90 \pm 0.04 \times 10^{33}$ ergs/sec

can be derived from long, independent series of observations.

An additional parameter, or series of numbers, is essential for the construction of an adequate theoretical model sun. These are the abundances of the chemical elements. For convenience, abundances are stated in terms of the abundance of hydrogen. In the sun, the relative abundances of the important constituents are

Hydrogen = 1.00	}	Q, the relative chemical abundance
Helium = 0.23		
Z (all other elements) = 0.02		

The values of L , M , R , Q , deduced from observation for the sun, are somewhere near the middle of ranges of these quantities as derived for other stars. It is, therefore, sometimes said that the sun is an average star, but for the stars inside an imaginary spherical shell surrounding the sun, with a radius so small that it is reasonably certain that all stars within the enclosed volume have been observed, a different interpretation is valid. Most of the stars close to the sun are cool, red, dwarf stars, probably the most numerous of all stellar varieties. Only ten of the fifty-five

nearest stars can be seen with the eye alone. Only three of these stars are brighter than the sun, most are exceedingly faint. Stars brighter than the sun are extremely rare, and the sun is outstanding in all of its properties.

Although the sun's L , M , R , and Q can be fitted into a coherent physical theory which, starting with a nuclear energy source (see SOLAR ENERGY SOURCES) near the sun's center, can trace the outward flow of energy from the deep interior, through complicated transformations, until a quantity and quality of radiation is predicted at the solar surface that agrees closely with observation; only the most primitive beginning has been made on a theory of the phenomena that occur in the observable regions of the sun: the photosphere, the chromosphere, and the corona.

The photosphere of the sun is the surface observed directly by the eye through a protective dense black glass screen. In very transparent skies with the help of special telescopes or at times of total solar eclipses, the chromosphere and corona may also be observed visually. All three of these layers show structures in widely different sizes and rates of change. The theoretical problems connected with the photosphere-chromosphere and the chromosphere-corona interfaces are nearly intractable, and the individual layers are understood only slightly better.

Nearly three centuries of telescopic observation of the solar photosphere define this part of the sun as that imaged on a photographic plate, or seen by the eye, using the light in a spectral band at least 1000\AA wide, centered near 5000\AA in the blue-green part of the spectrum. The portion of the sun thus recorded is a spherical shell, 5×10^8 cm thick, whose outer radius is the edge of the apparent solar disk, 7×10^{10} cm from the sun's center. Good photospheric photographs show a granulation composed of small (10^8 cm, or less) bright, circular, or hexagonal, structures all over the sun except extremely near the edge. These are the tops of convection cells that carry much of the solar energy. They are a few hundred degrees hotter than the five-thousand-plus degrees absolute that is consistent with the sun's assigned total luminosity. Occasionally, small dark spots (pores) appear among the granulations, and still more infrequently, a pore will develop into a larger dark complex, a sunspot region. Spot regions are some thousands of degrees cooler than the granulated photosphere in which they are immersed. They mark the locations of sizable magnetic fields, and the spot regions may cover some tenths of the sun's disk. Since the spot regions are nearly fixed on the solar surface, they may also be used as markers for the measurement of the rate of rotation of the sun.

The angular rate of rotation, deduced from sunspot, or spectroscopic, observations, varies from the sun's equator to its poles. It is greatest on the equator, there corresponding to a period of rotation of twenty-five days. At the pole the period of rotation is nearly thirty-five days.

Sunspots are the main sources of difficult problems of the photosphere. How can the

refrigeration of the spot regions be explained? What is the origin of the magnetic fields? How is the distribution of angular velocities established, and how is it maintained?

The chromosphere is most easily observed at times of total solar eclipse. It appears as red-purple narrow irregular ring just after the beginning and again just before the ending of totality. The average thickness of the chromospheric layer is 10^9 cm, but it is so tenuous that it is undetectable under conditions satisfactory for observation of the photosphere. For the daily observation of the chromosphere, a filter transmitting a spectral band not wider than 1\AA must be used to produce monochromatic solar images, and the center of the spectral band transmitted by the filter must be adjusted to coincide with the center of an emission line that appears strongly in the spectrum of the chromosphere. Nearly all observations of the chromosphere are made with light from the center of just two spectrum lines: the H-alpha line of hydrogen in the red part of the spectrum; and the K line of ionized calcium in the ultraviolet. The pictures show the chromosphere as a continually seething part of the solar atmosphere, subject to spectacularly sudden changes in the neighborhood of sunspots. The changes in the chromosphere and the dominating changes that occur in the underlying sunspots are considered together as the phenomenon of *solar activity*.

Long before regular observation of the chromosphere became possible, records of the numbers of sunspots had led to the discovery of a ten-year cycle of variation. As solar observation has become more nearly continuous, nearly every aspect of the sun's activity has revealed cyclical behavior closely synchronized with the variation in the numbers of spots. This is especially true for the changes in the chromosphere. The solar flares and extensive systems of solar prominences in the chromosphere are apparently organized and controlled by the magnetic fields rooted in the sunspot regions.

However, the smallest structures in the chromosphere are not obviously connected with sunspot activity. The smallest features in undisturbed solar areas far from the dominance of spot regions undergo damped oscillations that are nearly periodic with characteristic times about three hundred seconds. Perhaps these motions are enforced by the somewhat slower changes observed in the photospheric granulations on which the chromospheric structures are based.

Like the chromosphere, the corona can only be observed with the unaided eye at the time of a total solar eclipse, but with telescopic, spectroscopic, and other instrumental aids, it can be observed every day. It is the most extensive of the divisions of the sun's atmosphere, its outer limits lying somewhere beyond the distance of the earth from the sun where the corona becomes indistinguishable from the interplanetary medium. The corona changes both gradually and suddenly in rather sensitive connection with the spot regions. It is the principal source of the sun's x-rays, extreme ultraviolet, and radio emission,

and the combination of satellite observations near the short-wavelength limit of the solar spectrum and radio observations at the other end of the spectrum results in an attractive qualitative picture of solar activity.

At times of solar flare outbursts in the spot regions, observations in the two ends of the solar spectrum can be interpreted as indicating motion of an unknown disturbance, starting away from the sun with the beginning of the flare and moving outward through the chromosphere and the corona. Frequently, streams of particles associated with the flare activity reach earth, and are detected as cosmic rays and by their secondary effects such as terrestrial magnetic storms, but many of the changes induced in the corona by activity in spot regions seem to be connected with variations in local solar magnetic fields and not with streams of particles.

It should be evident from this brief synopsis of some of the general results of intensive observation of the sun for more than a century that solar theory is in a reasonably satisfactory state only for those parts of the sun that have not been observed. The problems for which theoretical help is badly needed to understand observed phenomena are derivatives of two main questions that are not necessarily unrelated: (1) What is the explanation of the observed variation of angular rotation from the sun's equator to its poles? (2) What are the physical bases of sunspot phenomena? When it is realized that not a single first-class spectrum of a sunspot exists, it is clear that remarkable opportunities in both observational and theoretical solar physics remain unexploited.

ORREN C. MOHLER

References

- Kuiper, G. P., Ed., "The Sun," Chicago, University of Chicago Press, 1953.
- Thomas, R. N., and Athay, R. G., "Physics of the Solar Chromosphere," New York, Interscience Publishers, 1961.
- Smith, H., and E., "Solar Flares," New York, The Macmillan Co., 1963.
- Gamow, G., "A Star Called the Sun," New York, Viking Press, 1964.
- Bray, R., and Loughhead, R., "Sunspots," London, Longmans, 1964.

Cross-references: ASTROMETRY, ASTROPHYSICS, SOLAR ENERGY SOURCES.

SOLID-STATE PHYSICS

Solid-state physics is the study of the crystallographic and electronic properties of solids, primarily of crystalline solids. It concerns itself with both the theoretical and experimental aspects of solids. Theorists attempt to apply both classical and quantum mechanical principles to the basic understanding of the nature of solids; experimentalists apply an enormous range of experimental techniques to the study of the properties

of solids. It is a broad field, merging with such associated disciplines as chemistry, metallurgy, ceramics, and electrical engineering; the boundaries between solid-state physics and these disciplines are not well defined.

The atoms of which a given solid is composed can be considered, for many purposes, to be hard balls which rest against each other in a regular repetitive pattern called the crystal structure. Crystal structures have a wide range of symmetries. Most elements have crystal structures of high symmetry; many compounds have complex, crystal structures of low symmetry. The determination of crystal structures, atom placement and the orientation dependence of numerous physical properties is an absorbing study, one which has occupied the lives of numerous geologists, mineralogists, physicists and other scientists for many years (see CRYSTALLOGRAPHY).

The hard-ball model of solids just described is too simple to explain many properties of solids. That solids can be deformed by external forces and that atoms in a solid possess vibrational energy imply deformability of the atoms, an attribute often built into the model by the assignment of springs to connect the atoms to their nearest neighbors. This ball-and-spring model has many successes, one important early use was that of Einstein to devise a reasonably successful quantum mechanical theory of specific heats.

The perfect crystal, one with all atoms on precisely defined lattice points, is nonexistent. By the way in which crystals are produced, either in a laboratory or by nature, defects in structure exist, often in profusion. These defects may be categorized by their geometry and size.

All crystals have atoms which occupy external surface sites and which do not have the correct number of nearest neighbors as a consequence. This surface is a seat of energy and is characterized by a SURFACE TENSION. Furthermore, internal surfaces exist, grain boundaries and twin boundaries, across which atoms are incorrectly positioned. In a crystal of reasonable size, say 1 cc, these two-dimensional defects, called *surface* defects, contain only about 1 atom of the solid in 10^6 , a rather small fraction.

Defects with extent in only one dimension are called *line* defects. The most prominent of these, the dislocation, is a line in the crystal along which the atoms have either an incorrect number of neighbors or neighbors which have not the correct distance or angle. In any cross section of a real crystal, one might expect to find a density of dislocations from 10^{11} to nearly zero per square centimeter.

Defects which have extent of only about an atomic diameter also exist in crystals—the *point* defects. Vacant lattice sites may occur; they are called vacancies. Extra atoms—interstitials—may be inserted between regular crystal atoms. Atoms of the wrong chemical species, impurities, may also occur.

The energy of the various defects is of concern to physics. The energy necessary to separate the

atoms of a solid into neutral, non-interacting atoms, the sublimation energy, is about 81,000 cal/mole for a typical solid, Cu, at room temperature. This is about 3.5 eV per atom. Surface energies, both of grain boundaries and free surfaces, are of order 1000 ergs/cm², about 2 to 10 eV per surface atom. Dislocation energies are of similar size per atom along the dislocation, 2 to 10 eV per atom, so the energy of a dislocation is some 10^{-3} to 10^{-2} ergs/cm of length. The energy necessary to produce a vacancy in Cu is about 1 eV, that necessary to produce an interstitial about 2 to 3 eV. Since energies of these magnitudes are much more than kT per atom at reasonable temperatures, defects can be produced only by conditions which exist during the manufacture of crystals, by external means such as plastic deformation or particle bombardment, or by decided fluctuations in thermal energy away from the average. The amount of energy which is bound up in defects is not large compared with the total thermal energy of a solid at ordinary temperatures. All the vacancies in equilibrium in Cu even at the melting point comprise less than 10 calories of energy per mole, much less than the enthalpy at 1357 K of more than 7000 cal/mole. Similarly, the total energy of the free surface of a compact block of 1 mole of Cu is less than 10^{-3} calorie.

Crystallographic defects do not remain stationary in the crystal, but they move about with time. Some of these movements serve to reduce the over-all free energy of the solid, others (these are chiefly movement of the point defects) may simply be the wandering of random walk. Since the defects move, in general, over a potential barrier large compared to kT when they move from one equilibrium site to another, their rate of motion is controlled by rather large fluctuations in energy. Consequently their rate of motion depends on temperature through an exponential factor $\exp(-\Delta H/RT)$, where ΔH is the enthalpy increase when the defect moves from the equilibrium site to the maximum in the barrier.

Defects are extremely important in controlling many properties of solids, in spite of the relatively small number of defects which exist in most solids. Such properties as the behavior of solids under mechanical stress; mass, heat, and charge transport; and a host of electrical phenomena are extremely sensitive to the presence of particular kinds of defects.

The electronic structure of solids is determined, in principle, solely by the electronic structure of the free atoms of which the solid is composed. Precise calculation of the electronic structure of a solid is, however, a difficult task, one that cannot be done for the general solid. Nevertheless, by using approximate models, many general features of the electronic structure can be deduced, especially when close interplay of theory and experiment is established. As for the crystalline structure of solids, two stages are useful in understanding the electronic structure. First, the perfect electronic structure is defined; then irregularities in this structure, termed defects, are

described. Although both the geometry and energy of crystalline defects are defined, description of the geometry of many of the electronic defects is not possible, and one must be content in such cases with description of the energy of the defect.

The electronic structure of the inner electrons of atoms in a solid is little different from that of the corresponding electrons in free atoms. The ENERGY LEVELS of the valence electrons are modified tremendously, however. The state functions of these outer electrons extend over the entire volume of the solid, and restrictions of quantum mechanics force the electrons to occupy a set of discrete but tightly packed energy levels. This *band* of levels has an extent of several electron volts. Importantly, unoccupied levels for the free atoms are also split into bands for the solid. The basic electronic nature of solids is determined by the position in energy of the lowest unoccupied levels relative to the highest occupied levels.

The solid is called a *metal* if excitation of electrons from the highest filled levels to the lowest unoccupied levels can occur with infinitesimal expenditure of energy. Metals have a high electrical conductivity, σ , and $d(1/\sigma)/dT$ is positive; they also have high thermal conductivity and are opaque to visible light. All of these properties are a result of the ease of excitation of electrons by external means.

Solids for which the lowest unoccupied levels lie several electron volts, say 2 or more, above the highest occupied levels are called *insulators*. They have low electrical and thermal conductivity (relative to the metals) and are transparent to visible light.

Semiconductors are solids which are intermediate in properties between the metals and the insulators. For them, the energy gap between filled and unfilled states lies in the range 0.1 to 1.5 eV.

In insulators and semiconductors, some electrons can be excited by thermal means from the occupied levels (the valence band) into the unoccupied levels (the conduction band) leaving behind an unoccupied state, called a hole, in the valence band. Such electron-hole pairs can also be produced by electromagnetic radiation of proper frequency. Both the excited electrons and the holes contribute to electrical conductivity in these crystals. Excited electrons and holes exist in metals too, but they are not necessary for conductivity in the metals.

Crystallographic defects, in general, are also electronic defects. In metals their most important role is to provide scattering centers for electrons. In semiconductors and insulators, however, the crystallographic defects can produce an almost limitless variety of conductive and optical phenomena. The crystallographic defects, virtually stationary in the crystal, provide sites in the crystal which have well-defined charge geometry in contrast to the highly mobile electron-hole pairs.

Just as crystallographic defects control many important properties which depend on mass transfer, so do electronic defects control many

important properties which depend on charge transfer. In fact, the entire technology of devices made of semiconductors and insulators is based on proper control and distribution of defects (including impurities).

MAGNETISM, also a proper subject of solid-state physics, is basically an inherent property of the perfect crystal. DIAMAGNETISM and PARAMAGNETISM, largely of academic interest, are little influenced by defects (except for such special properties as electron spin resonance). The basic properties of ferromagnets (Curie temperature, saturation magnetization and magnetostriction) are inherent properties of the perfect lattice. The technological application of ferromagnets, however, demands careful control of defects, which exert strong influence on such properties as hysteresis loss, permeability and coercive force (see FERROMAGNETISM).

Summarizing in a sentence: solid-state physics, in its broadest sense, is the study of the perfect and imperfect crystalline and electronic structures and properties of solids, ranging from attempts to understand these phenomena from the most fundamental point of view to the edge of the technological application of solids.

CHARLES A. WERT

References

- van Vlack, L., "Elements of Materials Science," Second edition, Addison Wesley Publishing Co., 1964.
 Wert, C., and Thomson, R., "Physics of Solids," New York, McGraw-Hill Book Co., 1964.
 Kittel, C., "Introduction to Solid State Physics," Second edition, New York, John Wiley & Sons, 1956.
 Seitz, F., and Turnbull, D., "Solid State Physics," New York, Academic Press, 1955-.

Cross-references: CRYSTALLOGRAPHY, DIAMAGNETISM, ENERGY LEVELS, FERROMAGNETISM, HEAT CAPACITY, MAGNETISM, PARAMAGNETISM, SEMICONDUCTORS, SOLID STATE THEORY, SURFACE TENSION.

SOLID-STATE THEORY

True solids possess long-range order not found in other phases of matter. The (approximate) periodicity of the potential in solids makes possible detailed investigations by powerful mathematical techniques which cannot be used in other condensed systems. Studies of phenomena related to the almost periodic potential comprise the bulk of modern solid-state theory. Research on gross deviations from periodicity, "imperfections," constitutes a sizable minority activity.

The one-electron problem in a truly periodic potential can be considered solved, in principle, with one reservation. Fast digital computers make the calculation of the "band structure," i.e., the dependence of electron energy on wave vector, $\epsilon(\mathbf{k})$, a matter of half an hour on the computing machine. (A few years, perhaps, is required to

develop a "program" capable of treating any crystal structure and potential.) The reservation refers to the uncertainty in knowledge of the potential to be furnished to the computing machine. Self-consistent methods have been constructed, but the treatment of "exchange" is not completely satisfactory.

Of course, electrons in a solid *do* interact, and a one-electron theory can not describe the correlation effects. Many-body theory is becoming increasingly important, and the collective behavior of the electrons is being investigated in various ways. Several types of cooperative phenomena are undergoing study.

The basic calculation of the dispersion of PHONONS (quanta of the vibrational field), i.e., the frequency of normal modes of oscillation vs their wave vectors, $\omega(\mathbf{q})$, is not as simple as for the electrons in a periodic potential. Not only are the interatomic forces difficult to specify, but there exist peculiarities, indeed singularities, in the density of states. Phonons also interact with each other, since the vibrational motion is not truly harmonic. Considerable activity is currently associated with these problems.

More difficult still are the questions which arise when the electron system and the phonons interact. Most theoretical work on pure systems is based on (1) a description of the electrons as if the potential were actually periodic, i.e., as if no phonons or even zero-point vibration were present, and (2) a description of the lattice vibration which ignores the electronic state of the crystal. For many purposes, this approximation is not as crude as it appears at first sight; if the electronic wave function actually is coherent over many atomic positions, and frequently it is, then the lattice vibration is indeed insensitive to small changes in the electronic state, and vice versa. Thus in a metal under ordinary conditions, one may ignore not only any change in the dispersion curve, but also any effect on the distribution function of phonons as the electronic state is changed, say by an electric field. In a semiconductor like Ge, the "phonon drag" problem may require that the electron and phonon distributions in the presence of a driving force be considered self-consistently, even though $\epsilon(\mathbf{k})$ and $\omega(\mathbf{q})$ may be treated as unchanged. Even this problem is not simple.

When imperfections are present, such as vacancies, impurities, dislocations, etc., it is necessary to consider the possibility of localized modes being created in the vicinity of these imperfections; that is, modes of lattice vibration of frequency ω may exist near the imperfections which could not exist in the perfect crystal. Furthermore there may occur a concentration of lattice modes of vibration in a certain frequency range, in the vicinity of the imperfection.

In addition, in some systems such as ionic crystals, it is necessary to go one step further, and consider "relaxation" of the lattice as an imperfection changes its electronic state. The motion of the neighboring atoms leads to a change in the electronic wave function, which in turn

may induce a change in the vibrational spectrum. Thus the simple Einstein relationship between absorption cross sections and emission transition probabilities must be modified, for example.

In the perfect crystal there exist simple selection rules based on momentum conservation which relate the initial and final values of the wave vector of the excited electron upon optical excitation. In the simplest case they are equal, except for the very small value of the wave vector of the incident photon. In many cases, however, there exists the possibility of "indirect" transitions, in which a phonon is simultaneously excited or absorbed. The probability of these transitions is reduced by a few orders of magnitude from that of direct transitions, but they are still observed and very important in macroscopic crystals. Theory has also developed to account for numerous new effects in optical properties of solids, such as nonlinear optical effects, laser action, 2-electron transitions, and 2-photon effects.

Magnetic behavior has received increasing attention in recent years. Nuclear and electron resonance phenomena have been powerful techniques for investigating the structure of solids (see MAGNETIC RESONANCE). Furthermore, the use of high magnetic fields has elucidated phenomena which were not capable of being investigated earlier. The theory of these phenomena has kept pace with them, if not preceded them. The theory of FERROMAGNETISM continues to occupy a major field of activity. The specific interaction among electrons which gives rise to the ferromagnetic behavior is not identifiable for a given system, although the general outlines of the theory are reasonably well understood.

The mechanical properties of solids are now understood at a far more fundamental level than they were a decade ago, through development of the theory of dislocations. Many new phenomena have been interpreted or predicted, such as spiral growth, "climb," dislocation networks, "whiskers," decoration with colloidal particles, etc. This has been a most active field of research (see SOLID-STATE PHYSICS).

In the theory of transport phenomena there are attempts at more sophisticated theory than the conventional one-electron Boltzmann equation will allow. Density matrix methods are applied, for example, and scattering probabilities are computed without making use of such approximations as spherical energy surfaces, Born approximation, etc., and the importance of Umklapp processes is now better appreciated.

Major changes in the theoretical techniques that have occurred recently are the use of fast computing machines, of course, as well as the increasingly widespread application of group theoretical methods. Green's functions and field theoretical methods in general also are appearing more and more in the literature.

The concept of quasi particles is now widespread, and we have names for helicons, excitons, magnons, phonons, polarons, plasmons, etc. Much modern theory is devoted to considerations

of the existence and the propagation of these quasi particles.

The existence of SUPERCONDUCTIVITY, which represented such a mystery for so long, has now been explained in terms of electron-phonon interaction. At present, people are trying to see if other interactions could likewise give rise to the existence of superconductivity. There is some speculation that high-temperature superconductors might exist.

Work on solid-state theory is widespread. Notable is the work in the United States, the Soviet Union, Japan, and England, but many other localities are contributing also.

DAVID L. DEXTER

References

Kittel, C., "Introduction to Solid State Physics," Second edition, New York, John Wiley & Sons, 1956.

Seitz, F., and Turnbull, D., Eds., "Solid State Physics," Vols. 1-15, New York, Academic Press, Inc.

Cross-references: ENERGY LEVELS, EXCITONS, FERMION SURFACE, FERROMAGNETISM, MAGNETIC RESONANCE, MAGNETISM, PHONONS, SEMICONDUCTORS, SOLID-STATE PHYSICS, SUPERCONDUCTIVITY, TUNNELING.

SONAR*

The term *SONAR* is a coined word derived from the phrase *SOund NAvigation and Ranging*. The term generally refers to the principles employed in the design and operation of systems that utilize acoustic energy transmitted in an ocean medium; while the systems themselves are referred to as sonar systems. Thus, sonar may be defined as a branch of applied acoustics concerned with the utilization of the ocean as the transmitting medium.

The problem of sonar is threefold: (a) understanding the transmission of acoustic energy through the transmitting medium, (b) developing sources which convert mechanical or electrical energy into acoustic energy, and (c) developing receivers which convert the acoustic energy back into mechanical or electrical energy.

Whenever a body vibrates in a fluid, longitudinal waves are formed, which propagate outward from the vibrating body. The particles of the fluid are set in motion, and temporary stresses are produced which increase and decrease during each vibration. The motion of the particles gives the fluid kinetic energy while the stresses induce potential energy. The sum of the two energies is called acoustic energy.

Traditionally, the starting point for a discussion of the transmission of acoustic energy in a fluid is to assume a point source radiating acoustic energy in an ideal homogeneous nonabsorptive

* The opinions and assertions contained herein are the private ones of the writer, and are not to be construed as official, or as reflecting the views of the Navy Department or the Naval Service at large.

medium of infinite extent. Under these assumptions, the energy from the source will radiate outwards with the wave front forming a spherical shell. As the radius of the shell increases, the sound intensity decreases. In practice it is customary to express the sound intensity by means of a logarithmic scale. The most generally used logarithmic scale is the decibel. The intensity level in decibels of a sound of intensity I is defined as $10 \log (I/I_0)$ where I_0 is a reference intensity. The intensity level can also be expressed as $20 \log (P/P_0)$ where P is the pressure and P_0 the reference pressure, usually 1 dyne/cm^2 in underwater acoustics. In this discussion the terms in all equations are expressed in decibels. The decrease in intensity as the shell increases in radius is called the spreading loss. The spreading loss from a unit range of R_0 to a range of R is $10 \log [(I_{\text{intensity at } R_0}) / (I_{\text{intensity at } R})] = 10 \log (R^2 / R_0^2) = 20 \log R$. In most applications the use of such a simple model has been inadequate.

A more realistic model considers the following factors: the water-earth interface (bottom), the water-atmosphere interface (surface), the absorption of acoustic energy in the medium, the presence of foreign material in the medium, and the distribution of sound velocity. Considered as an acoustic medium, the waters of the ocean form a thin layer on the earth's surface. Some of the acoustic energy radiated into this layer by a source will reach either the surface or the bottom. At either of these surfaces abrupt discontinuities in acoustic properties occur. Because of these discontinuities part of the intercepted energy is reflected, part may be transmitted across the interface, and part may be scattered within the medium. Since the transmission of an acoustic wave in water is accompanied by a compression and expansion of the medium, friction will occur between water molecules. This friction results in the conversion of some of the acoustic energy into thermal energy. In addition to this frictional, or viscous, loss there is another loss of energy in seawater related to the salts which continuously undergo chemical changes because of pressure fluctuations. Energy losses associated with both of these phenomena are called absorption losses. Due to the presence of foreign bodies in the volume of water, reflection and scattering is not limited to the surface and bottom boundaries. Foreign matter and biological content vary widely in size and acoustic characteristics. All ocean waters contain such bodies which modify the direction in which the acoustic energy is transmitted. In sufficient number they may also modify and increase the total absorption loss. The effect of variations in sound velocity is to bend the wave front in the direction of the lower velocity. This bending of the wave front is referred to as refraction. Both refraction and reflection can result in the guiding of acoustic energy in certain directions.

The factors above affect the propagation of acoustic energy in seawater in two different ways. The first results in a spreading loss already mentioned, and the second results in a loss

referred to as attenuation. Attenuation consists of both the scattering and absorption losses. The spreading and attenuation are related to the distance the acoustic energy travels in different ways. An important difference is that the spreading loss frequently is relatively independent of frequency while the attenuation is a function of frequency.

There are three basic types of sonar systems: direct listening systems, echo-ranging systems, and communication systems.

In direct listening the acoustic energy is radiated by the target, which is the primary source. The acoustic transmission is a one-way process. In their more elementary forms direct listening sonar systems may be nondirectional and only give a warning that a primary source is in the vicinity of the searching vehicle; or directional, and permit determination of the bearing of individual primary sources relative to the listening platform. They generally do not give range. Direct listening is limited by the magnitude of the signal when it reaches the receiving point and the magnitude of the interfering noise which tends to obscure its reception.

In echo ranging, the sonar system projects acoustic energy into the water with the expectation that this energy will strike a target and enough of the energy will be reflected back to the searching platform so that it can be recognized as a target echo. The primary source of acoustic energy is in the searching platform, with the target, upon reflection of the energy, becoming a secondary acoustic source. The transmission of the energy is a two-way process. Echo-ranging sonar systems permit a determination of the bearing of a silent target, and by timing the echo-signal transmission and by knowing the velocity of sound in seawater, a range may also be obtained. Echo ranging is limited by the relative magnitudes of the signal and of the locally generated interference. In some cases the sonar performance is limited by reverberation, which is the acoustic energy returning by reflectors other than the target of interest.

Acoustically, sonar communications systems are similar to direct listening systems in that they utilize a one-way transmission path. Instrumentally, they are similar to echo-ranging systems, one located at each of the two points between which communications is to be established. In these systems coded pulses or voice modulated signals are transmitted by one system and received by the other.

To hear a target by direct listening, it is necessary that the acoustic level of the target less the transmission loss along the acoustic path from the target to the listening equipment be equal to or greater than the level of the background noise. This may be expressed as $L - H \geq N$ where L is the source level of the target, H is the one-way transmission loss, and N is the noise level. The size of this inequality depends upon operator skill, signal processing, and method of presentation. It is called the signal excess, E . This inequality can be written as an equation where $E = L - H - N$. This equation is called the direct-listening sonar

equation for an omnidirectional listening hydrophone. When using directional hydrophones a factor called the directivity index must be added to the right-hand member of the equation. The source level, L , is a measure of the amount of acoustic energy put into the water by the target vehicle and is equal to $10 \log$ (sound intensity at unit distance from the source). The transmission loss, H , is the sum of the losses related to refraction, surface and bottom reflection, absorption, and scattering. The noise, N , results from unwanted acoustic energies arriving from many different sources and normally consists of thermal, ambient, and self noises.

To see a target by echo ranging it is necessary that the acoustic level of the primary source less twice the transmission loss along the acoustic path from source to target plus the target strength be equal to or greater than the noise. This may be expressed, for a nondirectional receiver against a noise background, as $L - 2H + T \geq N$ where T is the target strength, a function of the reflecting characteristics of the target. Against a reverberation background the inequality becomes $L - 2H + T \geq R$ where R is the reverberation level. As in the case of direct listening, the inequalities can be expressed in terms of the signal excess as $E_N = L - 2H + T - N$ or $E_R = L - 2H + T - R$ where E_N and E_R are the signal excesses for noise and reverberation. The noise, N , comes from own-ship's noise and target noise. Reverberation, R , is the energy that is returned from the outgoing acoustic energy to the receiving equipment after having been reflected from reflectors in the medium other than the target. Reverberation sources usually are backscattering from the surface, bottom, and foreign particles in the water.

ERNEST R. ANDERSON

References

- Albers, V. M., "Underwater Acoustics Handbook," Pennsylvania State University Press, 1960.
 Horton, J. W., "Fundamentals of Sonar," U.S. Naval Institute, 1959.
 Kinsler, L. E., and Frey, A. R., "Fundamentals of Acoustics," New York, John Wiley & Sons, Inc., 1962.
 Officer, C. B., "Introduction to the Theory of Sound Transmission," New York, McGraw-Hill Book Co., Inc., 1958.

Cross-references: ACOUSTICS, ELECTROACOUSTICS, NOISE, ULTRASONICS.

SPACE PHYSICS

Space physics is usually taken to be the physics of phenomena naturally occurring in space, but frequently space is not clearly defined. It normally includes at least the outer portions of planetary atmospheres (in which densities are low enough to permit satellite flight), the outer portion of the sun's atmosphere, and interplanetary space. It frequently also includes galactic space and intergalactic space. It is also frequently taken to

include planetary studies and the study of the earth's atmosphere at altitudes above those attainable by balloons. With such a broad coverage, there are many special areas of study included within the scope of space physics, and several of these are described separately; for example, see ASTROPHYSICS, COSMIC RAYS, IONOSPHERE, PLANETARY ATMOSPHERES, RADIATION BELTS, RADIO ASTRONOMY, and SOLAR PHYSICS.

Earth's Outer Atmosphere. The earth's atmosphere extends far out into space, something which would not necessarily be expected on the basis of atmospheric properties near the earth's surface. There are two factors involved in this great extension. First, above 100 km, the atmospheric temperature increases rapidly with altitude, causing an outward expansion of the atmosphere far beyond that which would have occurred had the temperature stayed within the bounds observed at the earth's surface. Second, above about 100 km, the atmosphere is sufficiently rarefied so that the different atmospheric constituents attain diffusive equilibrium distributions in the gravitational field; the lighter constituents then predominate at the higher altitudes and extend farther into space than would an atmosphere of more massive particles. This effect is enhanced by the dissociation of some molecular species into atoms. The pressure p at altitude h , in terms of the pressure p_0 at the earth's surface, is given by

$$p = p_0 \exp \left(- \int_0^h \frac{\bar{m}g}{kT} dh \right) \quad (1)$$

where \bar{m} is the average mass of the atmospheric particles, g is the acceleration of gravity, k is the Boltzmann constant, and T is the atmospheric temperature. It is clear from this expression that the high temperatures and low particle masses above 100-km altitude act to maintain pressures at still higher altitudes in excess of those that would exist if the temperature and molecular weight were constant with altitude. Where diffusion equilibrium prevails, Eq. (1) applies to each constituent separately, provided \bar{m} is replaced by the particle mass for the particular constituent under consideration and p_0 is the partial pressure at some reference altitude where h is taken to be equal to zero.

The composition of the atmosphere does not change much up to 100 km; there is a region of maximum concentration of ozone (still a very minor constituent) near 20 to 30 km (see PLANETARY ATMOSPHERES), the relative concentration of water vapor falls markedly from its average sea-level value up to 10 or 15 km, and the relative abundance of atomic oxygen begins to become appreciable on approaching 100 km, due to photodissociation of oxygen by ultraviolet sunlight. Above 100 km, atomic oxygen rapidly increases in importance, due to the combined influence of photodissociation and diffusive separation in the gravitational field; above 200 km, atomic oxygen is the principal atmospheric constituent for several hundred kilometers. However, helium is even lighter than atomic oxygen, so its concentration

falls less rapidly with altitude, and it finally replaces atomic oxygen as the principal atmospheric constituent above some altitude which varies with the sunspot cycle between 600 and 1500 km. At still higher altitudes, atomic hydrogen finally displaces helium as the principal constituent. The hydrogen extends many earth radii out into space and constitutes the telluric hydrogen corona, or geocorona.

The temperature of the upper atmosphere, and hence its density, varies with the intensity of solar ultraviolet radiation, and this in turn varies with the sunspot cycle and with solar activity in general. The solar radio-noise flux is a convenient index of solar activity, since it can be monitored at the earth's surface. The minimum nighttime temperature of the upper atmosphere above 300 km has been expressed in terms of the 27-day average of the solar radio-noise flux \bar{S} at 8-cm wavelength, as follows,¹

$$T = 280 + 4.6\bar{S}$$

This varies from about 600°K near the minimum of the sunspot cycle to about 1400°K near the maximum of the cycle. The maximum daytime temperature is about one-third larger than the nighttime minimum.

Magnetosphere. The magnetosphere is that region of space in which the geomagnetic field dominates the motion of charged particles. Near the surface of the earth, the geomagnetic field resembles that of a dipole; the best-fit dipole is off center about 440 km and is inclined about 11° to the earth's axis of rotation. Well out into space, the field is severely deformed so that it no longer resembles a dipole field. The most apparent deformation is that due to a plasma of charged particles flowing away from the sun, a flow generally referred to as the solar wind. When the kinetic energy of the directed flow of the solar wind exceeds the energy density of the geomagnetic field, the plasma displaces the field. The diamagnetic properties of the flowing plasma tend to compress and confine the geomagnetic field,² at least on the side facing the sun (there is conjecture that on the side away from the sun the geomagnetic field might be indefinitely extended). Calculations indicate that the surface of the magnetosphere facing the sun is roughly hemispherical, with dimples over the earth's magnetic poles. Observations by space probes indicate that the distance that the geomagnetic field extends towards the sun is about 10 earth radii.³

The magnetosphere is the region of space within which many geophysical phenomena are confined. The Van Allen radiation belt (see RADIATION BELTS) apparently extends out to the surface of the magnetosphere on the side facing the sun, although this is not so on the side away from the sun, where the magnetosphere has a tail extending at least as far as the moon's orbit and perhaps much farther. Any charged particles from the IONOSPHERE that have sufficient energy to escape from the earth's gravitational

field are constrained by the magnetic field to remain in the vicinity of the earth.

The solar plasma that compresses the geomagnetic field and limits the magnetosphere flows away from the sun at a velocity that might be described as hypersonic—the ordered velocity exceeds the average random thermal velocity of the particles. Although the medium is so rarefied that collisions are rare, the particles can interact with one another through the agency of magnetic fields contained within the plasma. As a result, a collisionless shock wave develops in the flow before it reaches the surface of the magnetosphere.⁴ Space probes have shown that the shock front lies about 4 earth radii beyond the surface of the magnetosphere in the direction of the sun.³

Auroras are luminosities in the upper atmosphere at high latitudes caused by energetic particles, mainly electrons, that flow from the outer magnetosphere into the atmosphere. The cause of this electron bombardment from the magnetosphere is a long-standing scientific mystery. There is an obvious relationship to solar activity, but the mechanism that relates these two phenomena remains unknown (see AURORA AND AIRGLOW).

Interplanetary Space. Up until about 1954, interplanetary space was thought to be essentially a good vacuum, devoid of interesting physical phenomena except for occasional sporadic events such as the ejection of gas clouds by the sun or the passage of comets. It is now recognized that interesting physical phenomena are always present. The solar corona expands continually, giving rise to a steady outstreaming into interplanetary space of ionized gas from the sun. As was mentioned above, this outflow of ionized gas is generally referred to as the solar wind.

The steady outflow of particulate matter from the sun was first recognized by Biermann in 1954 on the basis of the deflection of comet tails away from the sun—a deflection that was too great to be explained on the basis of light pressure. Biermann at first overestimated the strength of the outflow, because of an underestimate of the strength of the interaction between the solar plasma and the cometary plasma. The solar wind was first observed directly and measured with some accuracy in the spacecraft Mariner II. This showed that the concentration of solar material near the earth was of the order of 5 protons/cm³, along with a corresponding concentration of electrons, moving with a velocity of about 500 km/sec.⁷

Parker⁵ has given the most satisfactory explanation of the solar wind, describing it as a continuous hydrodynamic expansion of the solar corona, with a continuous heat input into the outflowing gas for a substantial distance near the sun. The flow can be compared to the flow of gas through a rocket nozzle, where, in the case of the sun, gravity plays the role in restricting the gas flow that the throat plays in the rocket nozzle. Parker has also shown that the outflowing, electrically conducting gas must pull the solar magnetic field out radially and

that the rotation of the sun must twist the radial pattern into a spiral. This provides a magnetic connection, or a guiding path, from the western portion of the sun to the earth for any cosmic radiation produced in solar flares, something that is confirmed by observation.

A surprising property of interplanetary space, as observed by space probes, is the irregularity of the magnetic field. Although the average orientation of the field agrees with the spiral pattern predicted by Parker, many irregularities are present.³ The magnetic field energy density is mainly due to the irregular fields, and it is approximately equal to the thermal energy of the solar wind particles, which is approximately 1 per cent of the energy of ordered flow. The real significance of these observations seems not to have been recognized at this writing.

A satisfying concept for the termination of the solar wind at some finite distance from the sun has been provided by Axford *et al.*¹ As the solar wind moves outward from the sun, its concentration falls according to an inverse square law, and the dynamic pressure that can be generated by stopping it falls accordingly. At some point, it must become so attenuated that its continued forward hypersonic flow will be stopped by the galactic magnetic field. At this point, there must be a shock front and a conversion of ordered energy of flow to disordered thermal energy. The heated gas beyond the shock front should cool mainly by charge exchange between the high-temperature protons from the solar wind and cool hydrogen atoms from galactic space. After charge exchange, the hydrogen atoms will carry the energy away. The cool proton gas left behind is less electrically conducting than was the hotter gas, and the magnetic field is gradually released from the proton gas, allowing the magnetic field lines to merge and the proton gas to drift out into galactic space.

The hydrogen atoms that are heated beyond the shock front have high enough velocities to penetrate far into the solar system, even to the vicinity of the earth's orbit, before becoming ionized by solar ultraviolet radiation. The hydrogen atoms within a few astronomical units from the sun scatter hydrogen Lyman-alpha radiation emitted by the sun, and this scattered radiation can be detected with instrumentation flown in rockets. The rocket observations can be used to determine the concentration of the high-velocity neutral hydrogen atoms in interplanetary space, and this in turn can be interpreted in terms of the distance from the sun to the shock front beyond which the hydrogen originates. Patterson *et al.*⁶ have shown that this interpretation indicates that the distance to the shock front is about 20 astronomical units (AU). The solar magnetic field spirals around about three times in this distance, and in the outer portion of the solar system, the magnetic field lines in the plane of the ecliptic are approximately circular, with the sun at the center.

Such a pattern of magnetic field in the solar system can be expected to produce significant

anisotropies in the cosmic radiation, whether of solar or galactic origin. This is most pronouncedly true when cosmic radiation is released from a point on the sun that is connected by a magnetic field line to the vicinity of the earth, in which case the cosmic radiation appears to approach the earth's magnetosphere from a point forty or fifty degrees to the west of the sun. A weak anisotropy is produced in the low-energy galactic cosmic radiation, and the anisotropy becomes smaller for the higher energy radiation.

FRANCIS S. JOHNSON

References

1. Axford, W. I., Dessler, A. J., and Gottlieb, B., "Termination of Solar Wind and Solar Magnetic Field," *Astrophys. J.*, **137**, 1268-1278 (1963).
2. Beard, D. B., "The Solar Wind Geomagnetic Field Boundary," *Rev. Geophys.*, **2**, 335-366 (1964).
3. Ness, S. F., Scarce, C. S., and Seek, J. B., "Initial Results of Imp I Magnetic Field Experiment," *J. Geophys. Res.*, **69**, 3531-3569, (1964).
4. Nicolet, M., "Solar Radio Flux and Temperature of the Upper Atmosphere," *J. Geophys. Res.*, **68**, 6121-6144 (1963).
5. Parker, E. N., "Interplanetary Dynamical Processes," pp. 272, New York, Interscience Publishers, 1963.
6. Patterson, T. N. L., Johnson, F. S., and Hanson, W. B., "The Distribution of Interplanetary Hydrogen," *Planetary Space Sci.*, **11**, 767-778 (1963).
7. Snyder, C. W., Neugebauer, M., and Rao, U. R., "The Solar Wind Velocity and Its Correlation with Cosmic Ray Variations and with Solar and Geomagnetic Activity," *J. Geophys. Res.*, **68**, 6361-6370 (1963).
8. Spreiter, J. R., and Jones, W. P., "On the Effect of a Weak Interplanetary Field on the Interaction between the Solar Wind and the Geomagnetic Field," *J. Geophys. Res.*, **68**, 3555-3565 (1963).

Cross-references: ASTROPHYSICS, AURORA AND AIRGLOW, COSMIC RAYS, IONOSPHERE, IONIZATION, PLANETARY ATMOSPHERES, RADIO ASTRONOMY, SOLAR PHYSICS.

SPECIFIC GRAVITY. See DENSITY AND SPECIFIC GRAVITY.

SPECTROSCOPY

Spectroscopy is the branch of science in which the interaction of energy between electromagnetic radiation and matter is investigated. It is one of the few branches of science which yield directly, accurate information concerning the nature of substances in their own environment; it has been until recently practically the only experimental method for obtaining extraterrestrial information. Its importance as a major experimental method in physics, chemistry, and the biological sciences, can be demonstrated by the wide range of problems which can be solved with its techniques.

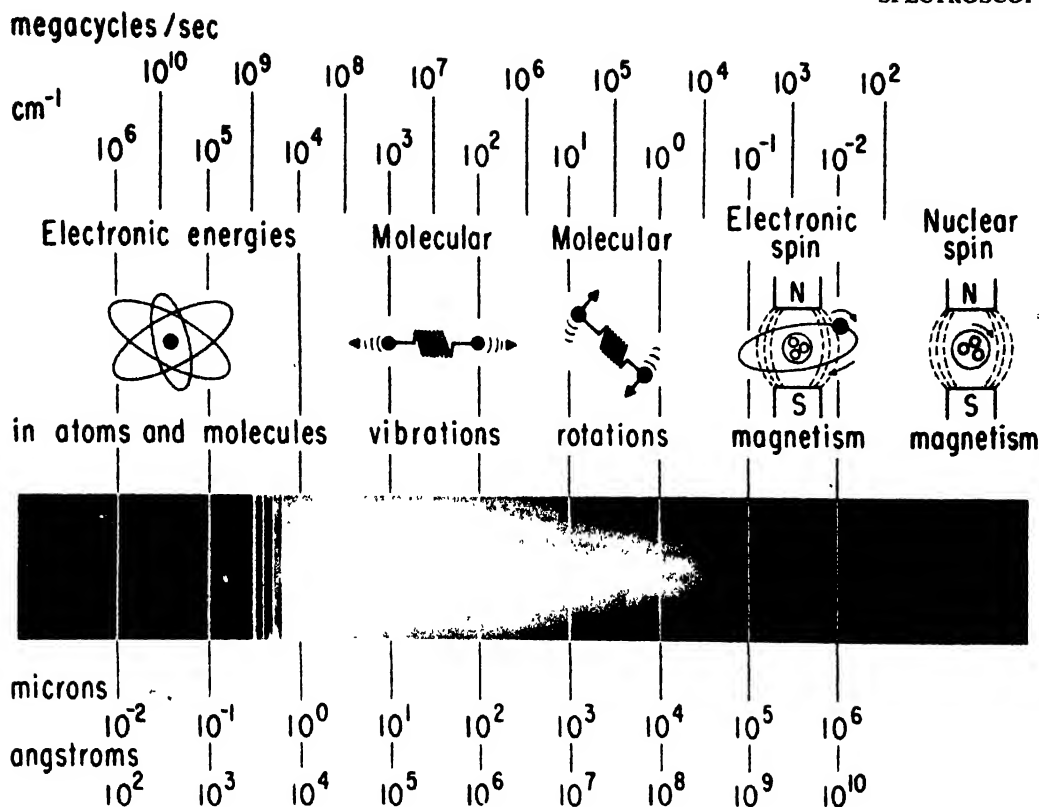


FIG. 1. Schematic representation of the electromagnetic spectrum in which the spectral regions and types of transitions are shown.

The broad field of spectroscopy can be subdivided in various ways, e.g., according to the energy (or frequency) range of the electromagnetic radiation studied or according to the nature of the transition involved. Several processes which result in absorption or emission of energy may occur within an atom or molecule, but fortunately each process is associated with a fairly definite frequency range, with very little overlap. Figure 1 is a schematic representation of the electromagnetic spectrum in which the various regions and the type of transitions which occur in each are indicated.

At very high energies or high frequencies, the transitions involve changes in the atomic nucleus and are independent of the environment of the nucleus. At slightly lower energy, inner shell electronic transitions occur. Irradiation of a sample with x-rays results in expulsion of electrons from the inner shells and then emission of x-rays as the electrons return to their normal states (x-ray fluorescence). The frequency of the x-rays emitted is independent of the state of chemical binding and depends only on the atom.

Irradiation with still lower-energy photons results in electronic transitions in the valence shell. The frequency of the radiation emitted when outer-shell electrons, which have been

excited either thermally or by electric arc, return to their ground states depends on the element involved and is usually independent of chemical state. The x-ray fluorescence mentioned above and the emission spectra in the ultraviolet and visible regions are very useful in identification of elements present in a sample. Atomic emission spectra are also useful in determination of electronic energy levels in atoms and in determination of composition and other properties of stars, planets and comets.

With the exceptions of flame spectroscopy and the emission from upper atmospheres, molecules are not usually sufficiently stable at the high temperatures required for thermal excitation of outer-shell electrons and, therefore, do not emit radiation in far ultraviolet, ultraviolet and visible regions. However, molecules do *absorb* radiation in the far ultraviolet, ultraviolet, and occasionally in the visible regions, as a result of the excitation of outer shell electrons. The absorption spectra in these regions are useful in studying the electronic states of small or unsaturated molecules and in the analysis of unsaturated organic compounds and inorganic complex ions. The energy absorbed by the molecules is usually rapidly converted to vibrational, rotational, or translational energy, but occasionally emission occurs (fluorescence and phosphorescence).

In the infrared region, changes in vibrational energy, usually accompanied by changes in rotational energy, are observed as absorption spectra. (Vibrational emission spectra have only recently begun to be studied in detail.) At still lower frequencies, in the far infrared and microwave regions, pure rotational transitions are observed.

Radiation in the microwave and radio-frequency ranges is used for studying the very low-energy transitions which result from reorientations of nuclear and electron spins in an applied magnetic field, i.e., for nuclear and electron magnetic resonance studies.

The occurrence of an energy change depends on the ability of the molecule or atom to interact with the electromagnetic radiation. In general, energy is absorbed only if the energy of the incident radiation is precisely the same as the amount required for the transition to take place.

Certain other criteria must also be satisfied. For example, the criterion for the absorption or emission of vibrational energy by a molecule in the infrared region is that a change in the electric dipole moment of the vibrating species must occur during the vibration. That is, not only must the energy of the incident radiation equal the energy of the vibration, but also the vibration must produce a temporary displacement of the electrical center of gravity. For a pure rotational energy change, the molecule must possess a permanent electric dipole moment. Nor do transitions take place between all different energy levels. The number of transitions allowed is limited by selection rules which can be justified theoretically on the basis of the limitations which must be introduced in order to obtain acceptable solutions for the wave equations. For example, pure rotational energy changes are limited to transitions between adjacent levels.

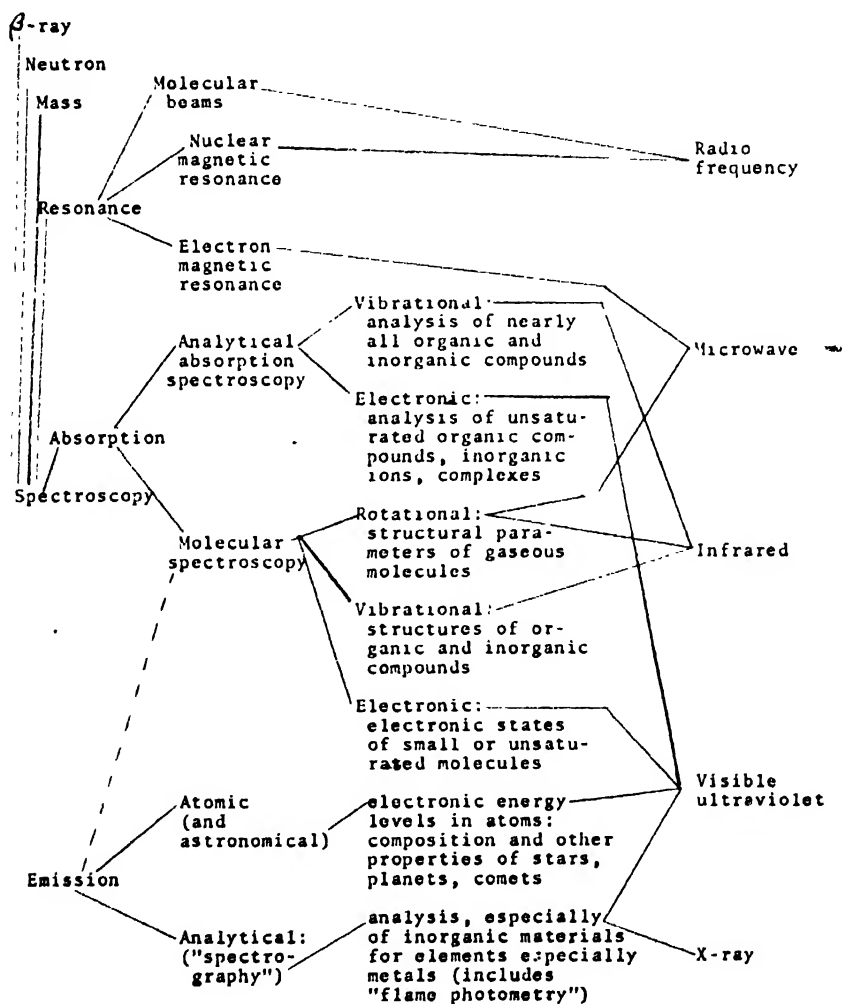


FIG. 2. Fields of spectroscopy (reproduced with permission from R. Bauman "Absorption Spectroscopy," New York, John Wiley & Sons, 1962).

The selection rules are occasionally defied, but the intensity of such a forbidden transition is usually low. Selection rules for linear molecules are fairly simple, but as the symmetry of the molecule decreases, the complexity of the selection rules increases.

Raman spectroscopy is an important exception to the rule that incident photons must possess exactly the correct amount of energy for the transition to occur. Under certain circumstances, frequencies in the visible or ultraviolet regions may be partially absorbed and cause the molecule to vibrate or rotate. The photon is then re-emitted with a new frequency, the Raman line, which is lower than the original and equal to the difference between the incident frequency and the vibrational frequency (see RAMAN EFFECT AND RAMAN SPECTROSCOPY).

The term "spectroscopy" has also been applied to some fields which do not involve electromagnetic radiation, but which use techniques similar to those of spectroscopy to separate a beam of particles according to their energy or other property. In β -ray spectroscopy, the energies of electrons emitted from nuclei are studied. In mass spectroscopy, charged particles are separated as a function of the ratio of charge

to mass. Figure 2 gives a summary of the fields of spectroscopy.

In order to indicate the broad scope of spectroscopy, some of the types of information which can be obtained and the experimental methods used to obtain them will be indicated. Studies in the ultraviolet, visible, infrared, microwave and radio-frequency regions have all found use in qualitative and quantitative analysis and in the determination of the geometry of simple molecules.

ATOMIC SPECTRA, which historically contributed extensively to the development of the theory of the structure of the atom and led to the discovery of the electron and nuclear spin, provide a method of measuring ionization potentials, a method for rapid and very sensitive qualitative analysis, a method for quantitative spectrochemical analysis, and data for the determination of the dissociation energy of a diatomic molecule. Information about the type of coupling of electron spin and orbital momenta in the atom can be obtained from the spectra obtained with an applied magnetic field. Other applications include the use of atomic spectra to obtain information about certain regions of interstellar space from the microwave frequency emission by hydrogen and the use of

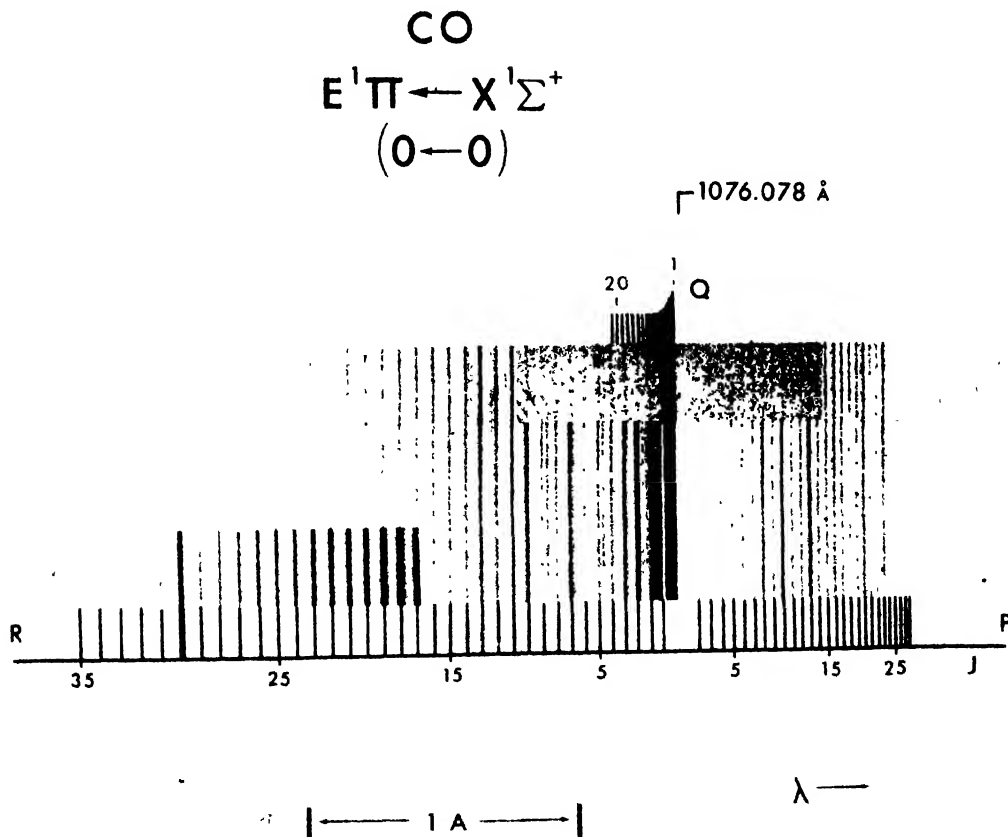


FIG. 3. High-resolution vacuum-ultraviolet absorption spectrum of the E singlet $\pi \leftarrow X$ singlet Σ^+ transition in CO. (Courtesy of S. G. Tilford, J. Vanderslice, and P. Wilkinson, U.S. Naval Research Laboratories).

atomic spectra to examine discharges in thermonuclear reactions.

Molecular electronic spectra arise from changes in electronic energy accompanied by changes in vibrational and rotational energy. Were only the electronic energy changed during a transition, as in atomic spectra, a single line would be observed. Instead, for each electronic transition a system of bands is observed, each of which is composed of a group of lines. These lines result from the changes in rotational energy in each pair of the vibrational energy levels associated with the upper and lower electronic states (see Fig. 3) and BAND SPECTROSCOPY).

Data obtained from vibrational analysis of band systems of diatomic molecules and radicals can be used to determine the anharmonicity of the vibrational motion, approximate values for the dissociation energy of the species studied, and force constants of the upper and lower electronic states, to calculate the value of α in the Morse equation, and to assist in the statistical calculation of thermodynamic functions. Measurement of the relative intensities of adjacent rotational lines in a band of an electronic transition can lead to the determination of the nuclear spin quantum number.

Electronic transitions in polyatomic molecules have been classified according to the type of orbital occupied in the upper and lower electronic energy levels. The position of the band and its sensitivity to the degree of conjugation in the molecule can be used to distinguish between the possible types of transition and hence to elucidate the nature of the bonding involved. Knowledge of the effects of conjugation and substitution on electronic spectra of organic molecules leads to the use of electronic spectra in identification of characteristic groups. Study of electronic absorption spectra is valuable not only in qualitative and quantitative analysis, but also in the investigation of steric hindrance, isomerism, and intermolecular interactions such as charge-transfer phenomena.

In the infrared region, considerable work has centered on associating a particular frequency with a characteristic group or structural unit for application in qualitative analysis. Internuclear distances, bond angles, spectral moments of inertia, and force constants are some of the more fundamental quantities which can be determined by infrared and Raman spectroscopy. These methods have also been useful in statistical calculation of thermodynamic quantities, such as entropy and heat capacity. They have proved very valuable in quantitative analysis, especially in cases for which chemical analysis is difficult or inapplicable. Figure 4 shows the changes in the infrared spectrum of a mixture of ethyl esters of maleic and fumaric acids on fractional distillation. The Raman spectrum of allene showing rotational structure is given in Fig. 5.

Spectroscopy in the microwave region includes the study of pure rotational changes and electron spin resonance phenomena. In addition to its qualitative and quantitative applications, micro-

wave spectroscopy provides probably the most accurate method of determining internuclear distances of linear, symmetric top and simple asymmetric top molecules and bond angles for the last two types of molecules. (In some of these determinations, the use of isotopes is necessary.) It provides a method of evaluating the quadrupole coupling constant and gives values of electric dipole moments with accuracy equal to or better than that obtained by the best of the dielectric constant methods. Microwave studies have also been made on the potential energy barriers opposing rotation in liquid and solid samples (see MICROWAVE SPECTROSCOPY).

The microwave region is almost always chosen for electron spin resonance studies, but some work has been done at radio frequencies. Some of the principal applications of the electron spin resonance technique are the determination of nuclear spin quantum numbers, elucidation of electronic states of ions in the iron group transition elements, and determination of the spectroscopic splitting factor (all of which involve study of crystalline solids, investigation of delocalization of electrons in molecules and among radicals in solution), and chemical analysis and identification of unstable free radicals.

The variety of problems for which nuclear magnetic resonance spectroscopy is applicable is considerable. Nuclear magnetic moments, internuclear distances (between nuclei whose nuclear spin quantum numbers are not zero) in crystalline solids, some molecular geometry, and quadrupole coupling constants have been evaluated.

Internal motion with molecules has been detected by NMR methods and the energy barriers evaluated. By means of chemical shifts and electron-coupled nuclear spin-spin interaction, structural determinations can be made. Chemical exchange studies in which the exchanged atom has a nuclear magnetic moment can be followed by NMR methods. Quantitative and qualitative analytical applications arise from the examination of the number, spacing, position and intensity of the lines in the spectrum. Figure 6 shows the spin-spin splitting of the three resonance lines in a high-resolution NMR spectrum of ethyl alcohol.

The unusual power of spectroscopic methods in scientific research is perhaps best emphasized by the diversity of their application in other fields—from measurement of intermolecular interaction forces, observation and study of properties of unstable species which cannot be captured, determination of temperatures and composition of astronomical bodies, to study of microorganisms.

ELLIS R. LIPPINCOTT
LINDA S. WHATLEY

References

- Harrison, George R., and Lord, Richard C., "Practical Spectroscopy," Second edition, New York, Prentice-Hall, Inc., 1965.

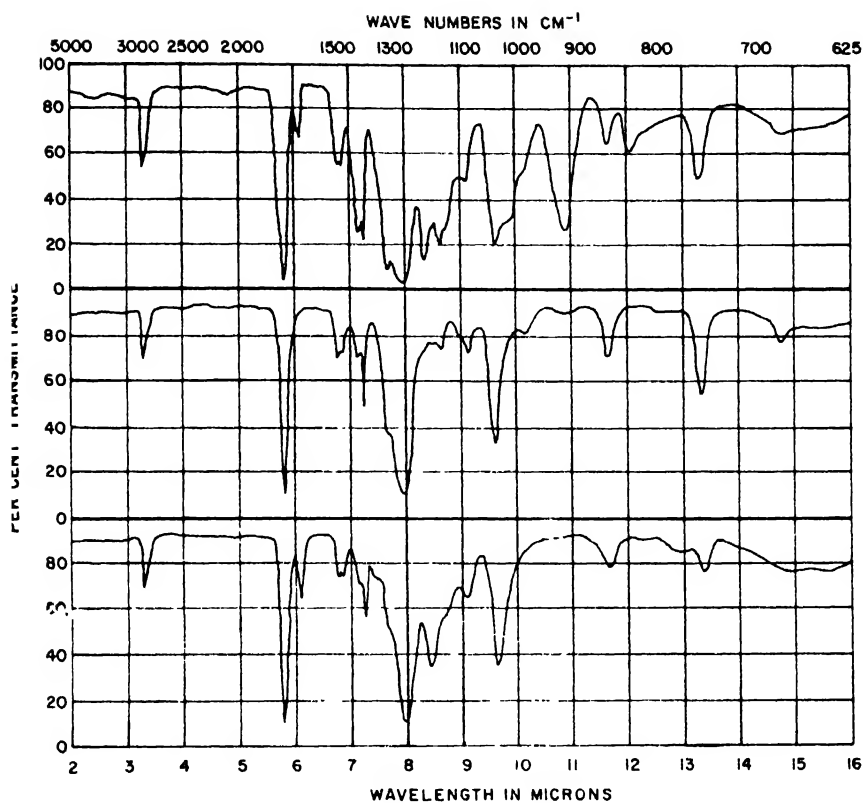


FIG. 4. The infrared spectra of successive distillation fractions of a mixture of the ethyl esters of maleic and fumaric acids.

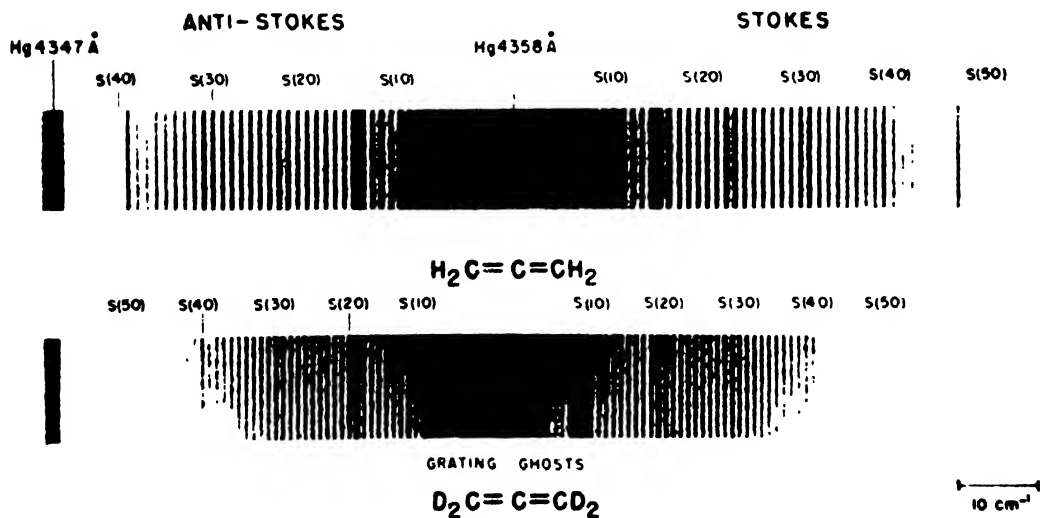


FIG. 5. The Raman spectrum of allene showing rotational structure.

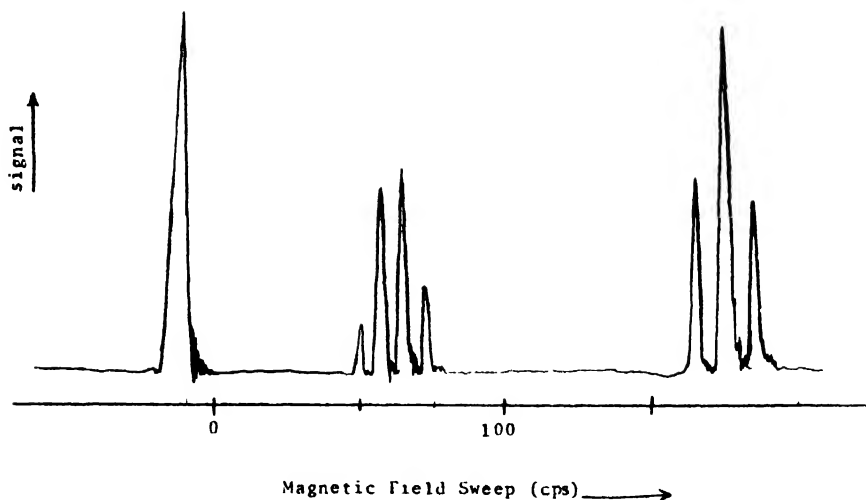


FIG. 6. High-resolution NMR spectrum of ethyl alcohol showing spin-spin splitting of the three resonance lines.

Herzberg, G., "Atomic Spectra and Atomic Structure," New York, Dover Publications, 1944.

Herzberg, G., "Spectra of Diatomic Molecules," New York, D. Van Nostrand, 1950.

Herzberg, G., "Infrared and Raman Spectra of Polyatomic Molecules," Princeton, N.J., New York, D. Van Nostrand, 1945.

Barrow, Gordon M., "Introduction to Molecular Spectroscopy," New York, McGraw-Hill Book Co., 1962.

Walker, S., and Straw, H., "Spectroscopy," Vol. I and II, New York, The Macmillan Co., 1962.

Townes, C. H., and Schawlow, A. L., "Microwave Spectroscopy," New York, McGraw-Hill Book Co., 1955.

Pake, George E., "Paramagnetic Resonance," New York, W. A. Benjamin, Inc., 1962.

Andrew, E. R., "Nuclear Magnetic Resonance," Cambridge Monographs on Physics, Cambridge, The University Press, 1958.

Pople, J. A., Schneider, W. G., and Bernstein, H. J., "High Resolution Nuclear Magnetic Resonance," New York, McGraw-Hill Book Co., 1959.

Clark, George L., Ed., "Encyclopedia of Spectroscopy," New York, Reinhold Publishing Corp., 1960.

Rao, C. N. R., "Ultraviolet and Visible Spectroscopy," London, Butterworths, 1961.

Cross-references: ABSORPTION SPECTRA; ATOMIC SPECTRA; BAND SPECTROSCOPY; ELECTRON SPIN; ENERGY LEVELS; LIGHT; MASS SPECTROMETRY; MICROWAVE SPECTROSCOPY; RAMAN EFFECT AND RAMAN SPECTROSCOPY.

STARK EFFECTS. See ZEEMAN AND STARK EFFECTS.

STATES OF MATTER

In writing about the "states of matter" it would seem reasonable to begin by deciding

what is meant by "matter" and by "state." Let, then, a short space be devoted to those questions.

As for "matter" consider it this way. Consider a brazen sphere, cube, and pyramid. What have they in common? They are brazen, or "of bronze." Now let the sphere be brazen, the cube golden, the pyramid of iron. What have they now in common? They are all metallic, or "of metal." Now let the sphere be brazen, the cube wooden, and the pyramid of earthenware. What have they in common? They are all solid, or "of solid stuff." Finally, let the sphere be brazen, the cube water, the pyramid outlined with smoke in the air. What have they in common? They are all material, or "of matter." So matter is that which is common to all material objects, but is not involved with immaterial things—the binomial theorem, for instance, or the right-angled isosceles triangle is not, as an abstract idea, material. This argument shows us, moreover, that we never have any experience of matter in the abstract. It is always "this ingot of iron," "that block of stone," "the pile of clay over there." However, just as, while we never see "man" but always John Jones or Sam Smith or some such individual, it is convenient to talk about the abstraction "man"; even so, the abstraction "matter" can be a convenient one.

As for "states," suppose we think this way. What is it that, above all else, distinguishes material objects? Is it not that they are tangible? For, on the one hand, a rainbow is visible, but we should not call it "material"; on the other hand, if we had something invisible, but which we could feel, we should declare it to be "material." So tangibility, it would seem, is the criterion. Now as to tangibility we might divide objects into: (a) unyielding, (b) yielding but quite tangible, and (c) hardly tangible. And these three classes, of course, we call solid, liquid, and vapour or gas: the three states of matter in which we are interested. This argument shows why we

distinguish matter primarily into these classes, and not, for instance, according to color. It is because it is above all else tangibility in which we are interested.

It is true, of course, that the division into three classes will leave some doubtful cases at the borders, just as does the division of living things into animal and vegetable. And we can always deal with a thing according to the aspect which predominates in the circumstances in which we are interested, so that we might think of pitch, say, as a solid at one time, a liquid at another.

Gases. A gas, of course, must be kept in some container, and it tends to expand the container, unless prevented from outside. We say that it exerts pressure; for instance, the pressure of air near sea level is about 14.7 psi, or just over 1×10^6 dynes/cm². It is found that the pressure of a given amount (by weight) of a gas in a container depends on the volume of the container, and on the temperature. The relation may be a very complicated one, but at sufficiently low pressure and sufficiently high temperatures, it is approximately true for all gases that

$$\frac{PV}{T} = \text{Constant}$$

P represents the pressure of the gas, V the volume and T the temperature measured from absolute zero, which is at 273 C or 460 F, approximately. This kind of equation is called an equation of state, and this particular one is the law of ideal gases, and it is said that a gas behaving thus is an ideal gas. Actually, of course, there is no such thing as an ideal gas, but any gas under suitable conditions will behave nearly ideally.

Liquids. It is sometimes said that there is no simple equation for liquids corresponding to the ideal law for gases. However, it is close to the truth, under many circumstances, to write

$$V = \text{Constant}$$

i.e., the volume of the liquid is constant independently of pressure and temperature. In fact, this is not true; the liquid is compressed a little by pressure, and the volume usually increases a little with increasing temperature. Nevertheless, this simple law is probably as close to the truth as is the law of ideal gases in many of the circumstances in which it is used.

Solids. Likewise a simple law can be given for solids, i.e.,

$$L = \text{Constant}$$

L being the length of the line joining any two points in the solid. That is to say, not only does the volume not change, but neither is the solid distorted by any action on it. Again, this is only an approximation to the truth, and probably about as valid.

Changes of State. We often tend to think of a given kind of stuff as being in a particular state; for instance we think of water as liquid, iron as

solid, and air as gas. However we know that almost anything may exist in any of the states; water, for instance may be liquid, solid, or vapor. The ways in which things change from one state into another can be very interesting. (In books on THERMODYNAMICS, it may be mentioned here, in which these things are discussed in great detail, the word "phase" is often used instead of "state" as used here.) In general, the division between the different states depends on the pressure and temperature of the sample. Increasing the temperature makes the material go to a "looser" state; e.g., if ice is heated somewhat, it melts, and if the water is heated more, it evaporates. Increasing the pressure tends to squeeze the material into a "denser" state, either liquid or solid according to circumstances.

The way in which the state of a given material, say water, varies with pressure and temperature can be shown well in a graph such as that in Fig. 1. Every point on the graph will represent a certain combination of pressure and temperature, in the way of analytic geometry. The graph is divided into regions marked "solid," "liquid" and "vapor." At temperatures and pressures such as T_1 and P_1 , which give a point in the region marked "solid," the material is solid. Likewise at temperatures and pressures such as T_2 and P_2 , giving a point in the "liquid" region, the material is liquid. And at temperatures and pressures such as T_4 and P_4 , it is a vapor. Thus, this graph indicates the whole behavior of the material with respect to temperature and pressure. Note that the line dividing the "liquid" from the "vapor" region just ends at a point at temperature T_c and pressure P_c . These are called the "critical point," and the "critical temperature" and "critical pressure", respectively. At temperatures higher than critical, there is no distinction between liquid

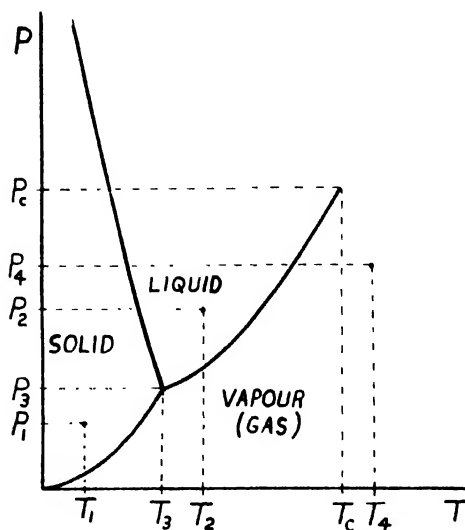


FIG. 1. Phase diagram: Plot of pressure vs temperature for a typical material.

and vapor. For water, the critical temperature is 374°C ; critical pressure, 219 atmospheres.

It can be seen that three parts of the curve meet as it were, at the point given by T_3 and P_3 . This point is called the "triple point." For water, the triple point is at about 0.01°C and 4.6 mm of mercury. Note that if you warm a solid at pressures lower than the pressure at the triple point, it will change to vapor without first melting to a liquid at all. Such "dry evaporation" is called "sublimation." For carbon dioxide, for instance, the pressure at the triple point is about 5.1 times atmospheric. Thus, solid carbon dioxide, i.e., "dry ice," when allowed to warm up at atmospheric pressure, sublimates directly into vapor, as is well known. A process the reverse of sublimation happens when frost forms on cold objects without any intervening stage of liquid.

Vapor Pressure. In Fig. 1, along the curve joining the triple point and critical point, the "liquid" and "vapor" regions meet; i.e., liquid and vapor can exist together. Note, however, that at a given temperature, they exist together only at a certain pressure. This pressure is called the vapor pressure of the liquid at the temperature concerned.

An analogous thing could be said about liquid and solid, or solid and vapor, along the appropriate curves.

Vapor Pressure and Relative Humidity. In a certain volume, say one cubic foot, of air, there is, along with the air, ordinarily a small amount of water vapor. Suppose that this same amount of water vapor were put into a vessel of the same volume, i.e., one cubic foot, which contained nothing else at all. There would be a small pressure, perhaps about $1/50$ of atmospheric pressure. This pressure is called the partial pressure of the water vapor.

The ratio of the partial pressure of water vapor in the air to the vapor pressure of water at the temperature of the air is called the relative humidity. (There is a bit of approximation here, because water vapor is not really an ideal gas.) For instance, at 70°F the vapor pressure of water is about 20 mm of mercury. (Standard atmospheric pressure is 760 mm of mercury.) If the actual partial pressure of water vapor were 10 mm the relative humidity would be 0.5 or 50 per cent.

Changes of Volume and Shape of Liquids and Solids. The "equations of state" proposed above for liquids and solids are really too drastic; they ordinarily expand when heated. Typically, the volume of a liquid increases by 0.05 per cent for each degree Fahrenheit increase in temperature. (Water may be anomalous in this, as well as other respects.) They are compressed by the application of pressure; typically, the volume is decreased by 0.01 per cent for each atmosphere of pressure.

Solids are generally compressed only about $1/50$ as much as liquids by the same pressure. It is more interesting to know how much force is needed to stretch a bar of solid material. For

steel, it takes about 30,000 pounds force for each square inch of cross section of the bar to stretch it by 0.1 per cent. For other solids, the force needed is mostly less, typically one-third to one-tenth of that.

The Surface of Liquids. A liquid, if it does not fill the container entirely, has an exposed surface. If an attempt is made to stretch the surface, for instance by withdrawing a horizontal piece of wire from the liquid, the surface acts as if it were an elastic sheet. The force necessary to stretch a unit length of surface is called the **SURFACE TENSION**. For instance, the surface tension of pure water around room temperature is about 72 dynes/cm. The addition of impurities, e.g., soap to water, reduces the surface tension greatly. The rising of liquids in fine tubes or in porous materials is connected with surface tension.

Molecular Theory of the States of Matter. The molecular theory considers all matter to be built up of molecules in some arrangement (or lack thereof!), each molecule, in turn, consisting of one or more atoms. The molecules attract one another with forces of electrical origin, and thus may stick together. However, they are also in motion or at least vibration, and this tends to make them break apart. The motion increases with increasing temperature, hence the structure "loosens." In a gas, around atmospheric pressure and room temperature, the molecules are about 3×10^{-7} cm apart; the forces between them are very small and they move almost freely and randomly. In so doing, they bombard the walls of the vessel; the effect of this bombardment is the pressure.

In a liquid or solid, the molecules are typically only about one-tenth as far apart. The forces then hold the molecules together; their motion is reduced to a vibration which never gets them far from one place. Occasionally, one molecule gets an extra hard push and flies away; this is evaporation or sublimation. According to the nature of the molecules, they may pile together as would a heap of spheres, or they may stick together in chains or other arrangements.

In a liquid the distance between molecules is about the same as in a solid. The different behavior may be because of a few "holes" in the liquid, i.e., places, a few per cent in all, where a molecule is missing. These spaces allow the liquid to flow, just as on a checker board when a few of the men are missing, the whole pattern can be moved by shifting men into and out of the holes. Of course this takes time, and it shows up in the viscosity or "slowness" of liquids. Molasses, for instance, is very viscous, water not very.

Various Kinds of Solids. In solids as was mentioned above, the molecules may be packed, arranged in chains, etc. Corresponding to these are the classes: metals, polymers, etc. Sometimes these are all spoken of as different states of matter, but from the present viewpoint it seems better to call them special kinds of solid.

References

General

Slater, J. C., "Introduction to Chemical Physics," New York, McGraw-Hill Book Co., 1939.

On the Forces which Hold Matter Together

Moelwyn-Hughes, A. E., "States of Matter," London, Oliver and Boyd, 1961.

On Thermodynamics

Pippard, A., "Elements of Classical Thermodynamics," Cambridge, Cambridge University Press, 1957.

Porter, A. W., "Thermodynamics," London, Methuen, and New York, John Wiley & Sons, 1951.

Cross-references: COMPRESSIBILITY; GAS; CONDENSATION; GAS LAWS; KINETIC THEORY; LIQUID STATE; PHASE RULE; SOLID-STATE PHYSICS; VAPOR PRESSURE AND EVAPORATION.

STATIC ELECTRICITY

Until it was accepted that galvanic currents are identical with moving electric charges, *static electricity* referred to all electric charge that was stationary or nearly so, including for example, that on Leyden jars and pyroelectric crystals. Nowadays the meaning is restricted to the more or less immobilized charge resulting from (1) redistribution of charge on a single body through *induction* by the charge on neighboring bodies; or (2) transfer of charge from one material to another by *contact* and subsequent separation. Sometimes the term is widened to cover the charge transferred through the action of external fields, as with electrostatic generators wherein charge is sprayed onto a conveyor by corona discharge; sometimes it is not, as with photoelectrets. The redistribution of charge by temperature gradients as with pyroelectrics and thermoelectrets, is generally excluded.

In scientific and technical work, it is often important to control static electricity, either to exploit its effects or to avoid them. For convenience we divide the problem into (a) generation of charge and (b) dissipation of charge. So far as generation is concerned, induced electrification is largely understandable in terms of classical physics; the fundamentals are well understood, and the applications are straightforwardly made as part of electrical engineering and safety engineering (see ELECTRICITY). Contact electrification, on the other hand, is far from being understood, and immediately comes to frontier problems in the modern theory of liquids and solids. So far as dissipation is concerned, the origin of the charge is largely irrelevant; a wide variety of physical and chemical processes of all degrees of complexity enter in analyzing the decay or neutralization of static charge.

Generation. Induction phenomena, as mentioned, are adequately described as a branch of classical electrostatics. Hence we limit our discussion here to the more poorly understood topic of contact phenomena. When two materials

differing at their surfaces in chemical composition, or even temperature or state of strain, are placed in contact, charge tends to flow from one to the other until their electrochemical potentials are identical. If the materials are subsequently separated, some portion of the transferred charge is retained, the potential difference between the materials increasing as the capacitance decreases. As a rule, this net transfer produces no noteworthy effects; in fact, a sensitive electrometer is ordinarily needed to measure it. With metal-metal contacts, the charge retained is invariably very small and is not detected by the senses. With metal-insulator or insulator-insulator contacts, the charge transferred while the materials remain in contact may be large, but it decreases during separation to an unimportant amount unless the insulator has quite high resistivity. At the speeds encountered in ordinary events, say about 100 cm/sec, the charge leaks back too fast to give noticeable effects whenever the resistivity of the more poorly conducting material falls below about 10^9 ohm-cm. At higher speeds, static effects become noticeable even at low resistivity, whereas at lower speeds they appear at very high resistivity.

The combination of moderate or high speed and high resistivity frequently produces large enough charge transfer that the surrounding medium breaks down electrically. In air at atmospheric pressure, the requisite charge densities on a uniformly charged plane conducting surface are about 8 esu (statcoulombs/cm²) or 2500 $\mu\text{C cm}^2$, to give a field of about 30 kV/cm just above the surface. This density represents about $2 \cdot 10^{10}$ electronic charges per square centimeter, so that only one in perhaps 10^9 surface atoms is charged even at the highest electrifications ordinarily occurring.

For practical purposes, the phenomena of interest in static electrification are the forces of attraction or repulsion resulting from excess charge (e.g., in textile processing, ore separating, and electrostatic copying), the occurrence of sparks and their consequences (e.g., in transfer of flammable liquids, or in processing of photographic films), and so on. From a fundamental point of view, these phenomena are simply consequences of transfer of charge and its subsequent behavior. Hence we may take as the central question in static electrification the following: Given two materials of specified chemical composition and physical state, what is the charge—in sign and amount—transferred when they are placed in contact under specified mechanical and ambient conditions, and then separated?

At the present stage of theory of matter, an answer is available only for some very special situations. For the rest of the cases, we must content ourselves with trying to get some guidance from crude models of the type to be described below. For brevity let us restrict our considerations to solid-solid contacts, even though liquid-solid contacts are almost as important technologically. Furthermore let us as a rule consider bodies that are uncharged before contact. We

assume first that a certain charge q_0 is transferred between two objects while they are touching. For many materials, q_0 will increase with duration of contact towards an equilibrium value q_∞ . We designate as α the degree of attainment of this value, writing $q_0 = \alpha q_\infty$. We should expect q_∞ to depend on the chemical nature of the materials, as well as on the mechanical nature of the contact, i.e., on the size and shape of the objects, and the normal force between them. Let us write q_∞ as a factor b ("b" for band structure, in the case of solids) dependent on the electrochemical properties of the materials, multiplied by a factor g ("g" for geometry) dependent on the mechanical parameters of the contact. Upon separation some of the charge returns to its origin, or is otherwise lost from the object; let us call f that fraction of the initial charge q_0 that remains to give the observed charge $q = q_0 f$. Upon combining all these relations, we get a four-factor formula

$$q = \alpha b g f$$

The utility of this expression can be assessed only by experience. Note that α and f are both positive numbers between zero and unity, and that g is intrinsically positive; the sign of the charge enters in b .

An important qualitative consequence of the above scheme is the existence of a *triboelectric series*, i.e., a listing of materials such that any one in it becomes positive when rubbed against another lower in the series. When the materials, the ambient conditions, and the mode of contact are reasonably well defined, such series are generally conceded to exist. An example is the following: Wool, nylon, viscose rayon (regenerated cellulose), cotton, silk, cellulose acetate, polymethylmethacrylate, polyvinylalcohol, polyethylene, polytetrafluoroethylene. In principle, one ought to be able to predict the position of a given substance from its chemical properties, but as yet the attempts have been more suggestive than successful.

So far as quantitative results are concerned, prediction of α and b is possible in principle when detailed information is available on the energy levels of the materials. As yet, only metal-metal systems have been found to be simple enough to be analyzed successfully. With insulators, surface state of high complexity may occur.

Prediction of g is hopefully made by estimating the area of "true contact," namely, the area where atomic fields interpenetrate. The effective width of this area should be proportional to some fractional power of the normal force, not less than $1/3$ and in some geometries approaching unity. The effective length of this area should again be proportional to a fractional power of the normal force in the case of stationary contact, but independent of normal force and proportional to the length of stroke in the case of sliding or rolling contact. Experiment tends to confirm these relations when the charge is not limited by atmospheric breakdown. In such cases a saturation value of charge is reached, its magnitude dependent on the atmospheric pressure.

Prediction of f is made by analyzing the time scale of the experiment. We expect f to be a function of the ratio of T , a time characteristic of the speed of separation, to τ , a relaxation time characteristic of the material of higher resistivity. The separation time T may be taken as l_0/v , where v is a speed of separation, and l_0 is some characteristic length; the relaxation time τ should be related to the time constant for redistribution of charge in a medium of dielectric constant ϵ and volume resistivity ρ . We have then $T/\tau = (l_0/v)/(\epsilon\rho)$, and we see that velocity may be traded for resistivity, since $f = f(T/\tau) = f(l_0/\epsilon\rho v)$. For ordinary materials at ordinary speeds, f is very small for ρ lower than 10^9 ohm-cm. For metal-metal contacts in particular, where $\rho \sim 10^{-5}$ ohm-cm, static effects disappear, their only remnant being the slight transfer due to contact potential difference.

Dissipation. Charge may be dissipated by currents within the body of the object or over its surface, or by currents within the medium in which the object is immersed. Charge in the interior of a body ultimately reaches the boundaries, decaying exponentially with a well-defined time constant $\epsilon\rho$. Charge on the surface of a body moves to attain an equilibrium distribution, decaying in a complicated fashion that is only approximately described by a time constant proportional to $\epsilon\rho$. Currents within the medium surrounding the object are described by the laws of electrical conduction in liquids or gases, as the case may be, in all their complexity.

Control of Static Electrification. Effects of static charge, beneficial or harmful, can be controlled by influencing either the production of charge or its neutralization.

Production of Charge. In the case of induced electrification, the fields between objects may be controlled by altering the potentials of neighboring conductors, usually by screening and grounding. In the case of contact electrification, the four-factor formula may be used as a basis for discussion:

CONTROL OF α : The only practical control is through fixing of the charge state of the object. (We must accordingly generalize our analysis to include objects initially charged.) If one object has lost so much charge that it can lose no more upon contact with another object, this second object will remain in its initial charge state. (As an illustration, a yarn running through an insulating or insulated guide may not pick up more charge after a brief initial period in which the guide electrically "saturates.")

CONTROL OF b : In principle, the electrochemical potential for one given material may be matched against that for some other material so that charge transfer between the two is negligible; in practice the potentials cannot be controlled closely enough, particularly over a range of ambient and mechanical conditions. More significantly objects usually must work against a variety of materials which will have a variety of electrochemical potentials. There have been proposed composite surfaces wherein a material

high in the triboelectric series is interspersed with one low in the series to give a small average electrification. The efficacy of such blending is high in some applications.

CONTROL OF g : Decrease of the normal force almost always produces a decrease in charge transfer. In cases where the dependence follows a fractional power law, it may not be useful to go to great lengths to decrease the normal force to very small values. With respect to path length, action is obvious though seldom practical.

CONTROL OF f : In practice, charge transfer is reduced by decreasing the product $\epsilon\rho v$. It is difficult to alter the dielectric constant by more than a small factor and hence to affect the rate of dissipation strongly in this way. The resistivity, on the contrary, may be changed greatly, either by changing the molecular structures of the body (say by grafting conducting segments onto polymers) or by adding conducting materials to the surface of the object (say by adding moisture or various antistatic agents). With hydrophilic materials, one traditionally alters the moisture content of the surrounding air, with that of the object following. The resistivity varies as a high negative power of the moisture content, and a change in relative humidity of a few per cent may bring about a tenfold difference in charge observed at a given time, or a tenfold difference in the time required to attain a given charge state. In a few special applications, the resistivity may be lowered by increasing the temperature. The factor r can be decreased by simply slowing down the process, of course, and also by changing the mode of contact, say from sliding to rolling (with concomitant change in g), or by lowering the relative speed of the two contacting objects.

Neutralization of Charge. Surplus charge inexorably is neutralized, since the resistivities of even the best nongaseous insulators seldom exceed 10^{18} ohm cm or so, and since natural radioactivity and cosmic rays produce mobile ions in surrounding fluids. When it is desired to hasten neutralization, charge must be supplied by other means. The conductivity of the body can be increased, especially at the surface, according to some of the ideas expressed earlier. More commonly, charge can be supplied through the surroundings. The medium in which the object is immersed can sometimes be rendered conductive by the electric field set up by the body itself, especially at surface regions of small radius of curvature, for example, at the points of tinsel or needles. Although the charge cannot be completely eliminated in this way, it can be made quite small if the ambient pressure can be brought near the minimum in the Paschen law for electrical breakdown in gases.

More often the surrounding atmosphere is ionized with the aid of external agents. Various commercial static eliminators have been developed to produce glow or spark discharges from high-tension wires or points. Electromagnetic radiation, in the form of x-rays or gamma rays, will ionize surrounding media, but it is often objectionable because of its hazards to personnel.

Particle radiation, in the form of alpha or beta rays, is very effective in producing ions, and is easier to control with respect to health hazard. Plutonium 240 with a half-life of 6600 years and americium 241 with a shorter half-life of 462 years are alpha emitters nearly free from gamma rays. They are modern substitutes for the nearly gamma-free polonium 210 with half-life of 138 days, and the gamma-active radium 226 with half-life of 1620 years. Flames produce copious ionization, but their action is usually only ancillary, as in flame-driers at the take-off of some printing presses.

D. J. MONTGOMERY

References

- Montgomery, D. J., "Static Electrification of Solids," *Solid State Phys.*, **9**, 139 (1959). Gives description of experimental techniques and results.
- Harper, W. R., "Electrification Following the Contact of Solids," *Contemp. Phys.*, **2**, 345 (1961). Develops another theoretical view.
- Gross, B., "Charge Storage in Solid Dielectrics," Amsterdam and New York, Elsevier Publishing Co., 1964. Gives an excellent selection of annotated references.

Cross-references: CONDUCTIVITY, ELECTRICAL; DIELECTRIC THEORY; ELECTRICITY; HIGH-VOLTAGE RESEARCH; VAN DE GRAAF ACCELERATORS.

STATICS

Statics is the branch of MECHANICS which studies the conditions of equilibrium of forces acting on particles or rigid bodies, or on inextensible cords, belts and chains. HYDROSTATICS, the study of the equilibrium of fluids, is usually not regarded as a part of statics in the conventional sense of the term.

Statics is the oldest branch of mechanics, some of its principles having been used by the ancient Egyptians and Babylonians in their constructions of temples and pyramids. As a science it was established by Archytas of Taras (ca. 380 B.C.) and primarily by Archimedes (287–212 B.C.); it was further developed by medieval writers on the "science of weights" such as Jordanus de Nemore (thirteenth century) and Blasius of Parma (fourteenth century). In the sixteenth century, it was revived by Leonardo da Vinci, Guido Ubaldo and especially by Simon Stevin (1548–1620) who laid the foundations of modern statics (inclined plane, equilibrium of pulleys, parallelogram of forces).

Although the laws of statics can in principle be derived from those of dynamics as a limiting case for vanishing velocities or accelerations, statics has been developed, since the end of the eighteenth century, independently of dynamics. Its fundamental notion, like that of dynamics, is the concept of *force*, representing the action of one body on another and characterized by its point of application, its magnitude and its direction (line

of action) or briefly by a VECTOR \mathbf{f} . Two equal and opposite forces whose lines of action are parallel and non-coinciding are said to form a *couple*. The *moment* or *torque* \mathbf{m}_O of a force \mathbf{f} about a point O is a vector whose magnitude is the product of the magnitude of \mathbf{f} and the length of the perpendicular distance of O from the line of action of \mathbf{f} , or in VECTOR notation $\mathbf{m}_O = \mathbf{r} \times \mathbf{f}$ (vector product), where \mathbf{r} denotes the vector from O to the point of application of \mathbf{f} .

The following four principles may serve as the basic postulates for statics. (1) *The principle of composition (addition) of forces*: two forces, \mathbf{f}_1 and \mathbf{f}_2 , with a common point of application A can be replaced by a third force, the resultant \mathbf{f} , which is obtained graphically (geometrically) as the diagonal, from A , in the parallelogram determined by the two given forces, or analytically (algebraically) as the vector whose components, usually with reference to a rectangular reference system, are the sum of the corresponding components of the two given forces, $f_x = f_{1x} + f_{2x}$, etc. (vector addition). The resultant of more than two forces is independent of the order of addition. (2) *The principle of transmissibility of force*: the point of application of a force acting on a rigid body can be transferred to any other point on the line of action of the force provided the point is rigidly connected with the body (sliding vector). (3) *The principle of equilibrium*: the necessary and sufficient condition for the EQUILIBRIUM, that is, absence of accelerated motion, of a particle is the vanishing of the resultant of all forces acting on the particle, or $\mathbf{F} = \sum \mathbf{f}_i = 0$. The condition for the equilibrium of a rigid body is the vanishing of the resultant of all forces as well as the vanishing of the resultant of their moments about an arbitrary point O , or $\mathbf{M}_O = \sum (\mathbf{r}_i \times \mathbf{f}_i) = 0$. If $\mathbf{F} = 0$, \mathbf{M}_O is independent of the choice of O . (4) *The principle of action and reaction* (Newton's third law): the force exerted by one body on another is equal and opposite to that exerted by the second body on the first and both forces lie along the same line of action.

These principles imply the following results. Two parallel forces can be added if additional compensating forces are introduced. Any set of coplanar forces, with the exception of couples, can be reduced to a single resultant. The sum of the moments of any two intersecting forces about any point in their plane equals the moment of their resultant about the same point (Varignon's Theorem). Any system of forces acting on a rigid body can be reduced, in an infinite number of ways, to a single resultant \mathbf{F} and a single couple \mathbf{M} . In all these reductions \mathbf{F} is uniquely determined but \mathbf{M} depends on the position of \mathbf{F} . There is one, and only one, line of action for \mathbf{F} , called *Poinsot's central axis of the system*, for which \mathbf{M} is parallel to \mathbf{F} . Hence every system of forces is equivalent to a *wrench* as this particular force-couple combination is called.

Statical analysis of framed structures or trusses, collections of straight members pinned or jointed together at the ends, is based on the preceding theorems. Such structures rest upon supports

whose reactions or pressures have usually to be determined in practical applications. The equilibrium conditions, according to (3), are two vector equations or, equivalently, six scalar equations and hence can be solved for no more than six unknowns (the reactions at supports and connections). If the reactions involve more than six unknowns, some of the reactions are statically indeterminate; if less, the body is said to be unstable. In case the unknown reactions arise from constraints, i.e., conditions restricting possible motions, a convenient method for the elimination of unknown reactions is the use of the *principle of virtual work* according to which the total virtual work (work due to a possible small displacement which need not necessarily take place) of the external forces acting on the body vanishes for any virtual displacement of the body. In general, internal forces, holding together the various parts of the structure, also have to be taken into account, e.g., for trusses which consist of straight members connected by joints.

In particular, parallel forces can always be replaced by a single resultant whose point of application is called the *center of the system of parallel forces*; it is invariant if all forces change their directions but remain parallel to each other. In this case the sum of the moments of these forces about any point equals the moment of their resultant about the same point; in particular, the sum of the moments about any point on the resultant is zero (generalization of the law of the lever). An important case of this kind is that of the earth's gravitational forces which act at a given place in practically parallel lines on every element of a not too voluminous body. The center of the system of forces, in this case, is the *center of gravity* and is identical with the center of mass or *centroid* which can easily be determined by summation for a system of discrete particles or by integration for continuous masses.

The study of friction, the resistance of a surface to the motion of a body upon it, belongs properly to applied mechanics. Since however many problems in statics involve, at least, considerations concerning *static friction* (the frictional force which just prevents motion), the study of FRICTION is often included in the science of statics.

MAX JAMMER

STATISTICAL MECHANICS

The object of statistical mechanics is the explanation of the macroscopical physical phenomena as consequences of the laws of motion of the atoms and molecules. Equivalently, statistical mechanics can be defined as the mechanics of systems of a very large number of degrees of freedom. Whereas in "ordinary" mechanics even the three-body problem cannot be solved exactly, statistical mechanics takes advantage of the large number N of degrees of freedom and tries to formulate exact *asymptotic* results in the limit $N \rightarrow \infty$ (in a certain well-defined way).

The fundamental laws of motion of the atoms and molecules are those of quantum mechanics; however, in many statistical mechanical problems, classical mechanics provides a sufficient approximation.

In ordinary classical mechanics, a problem is completely specified when the initial positions and momenta of all its particles are specified. Such information is impossible to obtain, and moreover is completely useless, for systems consisting of about 10^{24} particles. The only initial data which are interesting for such systems are of a macroscopic nature: density, local velocity, temperature at each point of a fluid, correlations between the density fluctuations in two points of the system, etc. There is a very large number of microscopic initial configurations of the molecules which is compatible with a given macroscopic specification. Hence, in order to describe such systems, Gibbs introduced the concept of an *ensemble*. This is defined as a set of a very large number of systems, all dynamically identical with the system under consideration (i.e., having the same hamiltonian H), differing in the initial conditions of the molecules but compatible with the macroscopic specification of the system.

The natural mathematical framework of such a description is the *phase space*, a many-dimensional space whose coordinates are the positions x_1, \dots, x_N (shortly: x) and the momenta p_1, \dots, p_N (shortly: p) of all the particles of the system. A point in phase space therefore represents a complete dynamical system in a definite microscopic configuration. A Gibbs ensemble corresponds to a cloud of points in phase space, which can usually be considered as a continuous distribution. The basic concept is therefore the *distribution function* in phase space $\rho(x, p; t)$, giving the density of the ensemble as a function of the positions and momenta of the particles (i.e., of the coordinates of the phase space) at time t . The connection between microscopic and macroscopic physics is then given by the following assumption: The observable value of a dynamical property of the system, (e.g., density, local velocity, average energy, etc.), $\bar{A}(t)$, is the average value of the corresponding microscopic dynamical function $A(x, p)$, weighted by the distribution function:

$$\bar{A}(t) = \int dx \int dp A(x, p) \rho(x, p; t) \quad (1)$$

Practically all functions $A(x, p)$ of physical interest are sums of functions involving only one or two particles. Hence it follows from Eq. (1) that the functions of real importance are the integrals of $\rho(x, p; t)$ over all but one or two particles: these are called reduced (one- or two-body) distribution functions. Their main importance comes from the fact that they remain finite in the limit $N \rightarrow \infty$.

The evolution of the system in time is described by the change in time of the distribution function. According to the laws of classical mechanics, the latter obeys a partial differential equation called the *Liouville equation*:

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^N \left[\frac{\partial H}{\partial p_i} \frac{\partial \rho}{\partial x_i} - \frac{\partial H}{\partial x_i} \frac{\partial \rho}{\partial p_i} \right] \equiv \frac{\partial \rho}{\partial t} + [\rho, H] = 0 \quad (2)$$

The bracketted expression $[\rho, H]$ is called the Poisson bracket of the hamiltonian H and the distribution function. The purpose of classical statistical mechanics is the solution of the Liouville equation in the limit $N \rightarrow \infty$.

In quantum statistical mechanics, the conceptual situation is much the same, but it is more complicated because of the proper statistical character of the quantum description of even a single system. Indeed, due to the HEISENBERG UNCERTAINTY PRINCIPLE, the momentum and the position of a particle can never be measured simultaneously with arbitrary accuracy. Hence the concept of a phase space has no meaning in quantum mechanics. The maximum information which can be obtained about a single system (in a "pure" state) is contained in its wave function $\Psi(x; t)$. The observable value of a dynamical variable in such a state is given in terms of the corresponding operator \hat{A} by the expression

$$\bar{A}(t) = \int dx \Psi^*(x; t) \hat{A} \Psi(x, t) \quad (3)$$

To this statistical aspect of quantum mechanics is added the proper indeterminism of statistical mechanics. Suppose the wave function of a single system (n) can be expanded in a series of orthonormal functions $\varphi_i(x)$ [see WAVE MECHANICS]

$$\Psi^{(n)}(x; t) = \sum_i a_i^{(n)}(t) \varphi_i(x) \quad (4)$$

In the statistical mechanical description, the single system is replaced by an ensemble, in which each system (n) is weighted by a density p_n . The role of the classical distribution function is now played by the *density operator* ρ introduced by J. von Neumann (1932). The matrix elements of this operator in the present representation are defined by

$$\rho_{ij} = \sum_n p_n c_j^{(n)} c_i^{(n)*} \quad (5)$$

the sum over n running over all the systems of the ensemble.

The averaging prescription which replaces Eqs. (1) and (3) is now:

$$\bar{A}(t) = \text{Trace } \rho \hat{A} \quad (6)$$

It is easily seen that this rule embodies the double averaging necessary for quantum and statistical mechanics. The evolution in time of the density operator is given by von Neumann's equation:

$$\frac{\partial \rho}{\partial t} + \frac{1}{i\hbar} [\rho H - H \rho] \equiv \frac{\partial \rho}{\partial t} + \frac{1}{i\hbar} [\rho, H] = 0 \quad (7)$$

H again being the (quantum) hamiltonian of the system. The classical Poisson bracket has been replaced by $(i\hbar)^{-1}$ times the commutator of the operators ρ and H .

Quantum statistical mechanics shares with the usual quantum mechanical many-body problem the following characteristic feature [see WAVE MECHANICS]. It is known that in a system of several identical particles, the latter are indistinguishable. In order to satisfy this requirement, the wave function of the system must be either symmetrical or antisymmetrical with respect to a permutation of any two particles. This symmetry requirement introduces the classification of quantum statistical systems into systems of *bosons* (Bose-Einstein statistics, symmetric wave functions) and systems of *fermions* (Fermi-Dirac statistics, antisymmetric wave functions). The purpose of quantum statistical mechanics is the solution of Eq. (7) with the proper symmetry condition.

The simplest solutions of Eqs. (2) or (7) are the time-independent solutions: these are functions of the hamiltonian H alone. They describe systems in equilibrium. A particular function of the hamiltonian is the *canonical distribution*:

$$\rho = Ce^{-H/kT} \quad (8)$$

where C is a normalization constant and k is called the Boltzmann constant. It can be shown that this distribution represents a system in thermodynamical equilibrium at temperature T . The main result in classical equilibrium statistical mechanics is the following: Consider the function

$$Z(V, T) = (N! h^{3N})^{-1} \int dp \int dx e^{-H(p, x)/kT} \quad (9)$$

the integration extending over the whole volume occupied by the system in phase space. $Z(V, T)$, as a function of the volume and of the temperature T , is called the *partition function*. It can be shown that the Helmholtz free energy $F(V, T)$ of the system [see THERMODYNAMICS] is given by

$$F(V, T) = -kT \ln Z(V, T) \quad (10)$$

The importance of this formula lies in the fact that the knowledge of a thermodynamic potential such as $F(V, T)$ enables one to calculate all thermodynamic properties (pressure, entropy, specific heat, etc.) by simple differential operations. A completely analogous result holds in quantum statistical mechanics.

Although the basic problem of equilibrium statistical mechanics is solved in principle (i.e., it is reduced to quadratures), the explicit evaluation of the $6N$ -fold integral occurring in the partition function poses a formidable mathematical problem. Only in the case of systems of noninteracting degrees of freedom can one calculate the partition function exactly. Moreover, essentially three main groups of systems of interacting particles are thoroughly understood at the present time. All of these systems are characterized by the following hamiltonian:

$$H = H_0 + \lambda V \quad (11)$$

where H_0 is the hamiltonian of a set of independent particles, whereas V describes the interactions and λ characterizes the size of the interactions. The three cases mentioned are the following:

(a) *Weakly coupled gases*. These are systems in which the interparticle interactions are weak; the partition function can then be expanded in a power series in λ .

(b) *Dilute gases*. In real gases, the particles interact through forces which have a very short range (the molecules can usually be idealized as hard spheres). In a dilute gas the molecules move most of the time in straight lines and occasionally suffer collisions, which are more and more frequent and involve more and more particles simultaneously as the density of the gas increases. H. D. Ursell (1927) and J. E. Mayer (1937) have shown that the partition function, and hence the pressure, can be expanded as a power series in the density; the result is the famous *virial equation of state* (or *cluster expansion*). The original derivation of this equation made extensive use of a diagram technique, a procedure which establishes a one-to-one correspondence between certain mathematical expressions and certain graphs; such diagram techniques proved to be of major importance in modern perturbation theory. The coefficients of the various powers of the density (or virial coefficients) are expressed in terms of so-called cluster integrals, describing correlations of various types between a given number of particles (see KINETIC THEORY).

The cluster expansion breaks down as the density increases and reaches the critical point. The condensation phenomena and the LIQUID STATE are not yet clearly understood.

(c) *Plasmas*. The ionized gases, or plasmas, have a radically different behavior as compared to "usual" gases. This is due to the long range of the Coulomb forces. As a result, one cannot speak of collisions in the ordinary sense: the interactions have a markedly collective character, involving many particles simultaneously. Mathematically, the various virial coefficients turn out to be divergent. J. E. Mayer (1950) has shown that by rearranging the cluster expansion and by performing summations of certain subseries, one can obtain a convergent equation of state for a plasma, the first term of which agrees with the one calculated by Debye and Hückel in their famous theory of electrolytes (which is semi-phenomenological) (see PLASMAS).

Whereas equilibrium statistical mechanics is in a state where the difficulties are mainly mathematical, nonequilibrium statistical mechanics is still in a state where the principles and the ideas are not yet completely clarified and unified. Most of our present knowledge has been achieved in the last twenty years, and this field is still in rapid growth.

The main problem here is to understand the basic paradox of irreversibility: whereas the laws of mechanics are invariant with respect to an inversion of time, the macroscopic evolution (e.g., heat conduction, dissipative phenomena, etc.) is irreversible. In older theories (Boltzmann), it was argued that due to the large number of particles and their complicated motions, one could invoke a probabilistic argument which, superposed to mechanics, yields readily an explanation of irreversibility.

There exist at present a few formalisms for the treatment of nonequilibrium mechanics (they are all equivalent). The basic idea in the modern theories is to avoid the probability arguments. Irreversibility appears as a property of the solutions of the mechanical equations, for systems of many particles and over sufficiently long times. More exactly, it can be shown that the Liouville equation reduces, for large systems, to a rather different equation, called the *master equation*. The latter describes a behavior which has a mixed character of reversibility and of irreversibility. In the three limiting cases mentioned above (weak coupling limit, dilute gas limit, plasma limit), there appear two natural time scales: the duration of a collision and the time between two collisions. The latter is much longer than the former. It can then be shown that for times much longer than the short time scale, the *reduced* distribution functions evolve practically independently of the rest of the correlations. They obey *kinetic equations* which describe an irreversible approach to equilibrium. In most cases other than the three limits described, the two time scales are no longer widely separated, and the evolution is much more complex.

One of the practical purposes of nonequilibrium statistical mechanics is the calculation of transport coefficients (thermal and electrical conductivity, viscosity, etc.) from molecular data. This can be done by starting from the kinetic equations. Alternatively, it has been shown recently (M. S. Green, R. Kubo) that a compact expression for most transport coefficients can be obtained from general arguments. The "Kubo formulas" express these coefficients in terms of an autocorrelation function of two microscopic currents (electrical currents, heat flow, etc.) averaged over the equilibrium distribution function. These two equivalent approaches can be combined in order to form a basis for the rigorous calculation of transport coefficients (see TRANSPORT THEORY).

Many other important results have been obtained recently in statistical mechanics; they cannot be reviewed in such a short article. The interested reader is referred to the existing textbooks, such as those given below.

RADU C. BALESCU

References

- Huang, K., "Statistical Mechanics," New York, J. Wiley & Sons, 1963.
- de Boer, J., and Uhlenbeck, G. E., Eds., "Studies in Statistical Mechanics," Vol. 1, New York, Interscience Publishers, 1962.
- Prigogine, I., "Non Equilibrium Statistical Mechanics," New York, Interscience, 1963.
- Balescu, R., "Statistical Mechanics of Charged Particles," New York, Interscience Publishers 1963.

Cross-references: BOLTZMANN'S DISTRIBUTION LAW, BOSE-EINSTEIN STATISTICS AND BOSONS, FERMI-DIRAC

STATISTICS AND FERMIONS, KINETIC THEORY, HEISENBERG UNCERTAINTY PRINCIPLE, LIQUID STATE, PLASMAS, THERMODYNAMICS, TRANSPORT THEORY, WAVE MECHANICS.

STRONG INTERACTIONS

Only two types of fundamental interaction operate in classical physics, which account for all macroscopic and chemical phenomena including atomic structure. There are the gravitational forces which control the motion of bulk matter, which is electrically neutral—particularly the motion of the planets in the solar system. Then there are the electromagnetic forces which govern the motion of the electrons in atoms, and also give rise to the interactions between atoms and, hence, form the basis of all chemical reactions.

An atom has a radius of about 10^{-8} cm. In an atom, a cloud of electrons orbit about a central nucleus with a radius of about 10^{-12} cm. These nuclei are made up of nucleons—protons and neutrons. The total charge, which is determined by the number of protons, fixes the chemical nature of the atom. The nuclei are extremely stable, remaining unchanged through the most violent chemical reactions. The gravitational forces within a nucleus are completely negligible compared with the electrical repulsion between the charges on the tightly packed protons. Since the electrical forces tend to blow the nucleus apart, it follows that a completely new, specifically nuclear, force must be operating to form these stable structures. This force is of short range, only effective when the nucleons are less than 10^{-12} cm apart. Within this range, it is strong enough to overcome the powerful electrical repulsions. It is known as the *strong nuclear interaction*, to distinguish it from the other specifically nuclear interaction—the *weak* interaction—which causes the spontaneous disintegration of subnuclear particles (see WEAK INTERACTIONS). It is the strong nucleon interaction which is the source of energy in nuclear weapons and in nuclear reactors.

To investigate the workings of the strong nuclear interaction a beam of high-energy protons from a proton accelerator is directed at a target of liquid hydrogen (more protons) and scattered. At low energies, below 300 MeV in the laboratory, the protons are merely deflected by the strong nucleon-nucleon interaction. Above this energy, new particles are produced in the collision. The first to appear are the π -mesons, or pions. As the energy increases, more and more particles are found. The least massive are stable as far as the strong interaction is concerned, but they disintegrate through the weak interaction with mean lifetimes of about 10^{-10} second. The more massive particles decay via the strong interaction into other strongly interacting particles in about 10^{-23} second. These are tabulated in the article on ELEMENTARY PARTICLES.

Because of the strength of this interaction, it has not been possible to discover the details of the

corresponding dynamics. However, considerable progress has been made by studying the conservation laws which govern the collisions of strongly interacting particles.

The subnuclear particles are specified according to their properties which correspond to those physical quantities which are conserved (i.e., do not change) in any collision which is dominated by the strong interaction. These are typically the mass, which contributes to the energy, and the spin which is part of the total angular momentum. If a particle has spin J (in units of $\hbar/2\pi$) there are $2J+1$ possible states, corresponding to the orientations of the spin axis allowed by quantum mechanics. These states are distinguished by the magnitude of the component J_z which ranges by integers from $+J$ to $-J$.

Other conserved quantities are the electric charge, the so-called hypercharge, and the baryon number. These are also tabulated in the article ELEMENTARY PARTICLES.

These subnuclear particles appear in multiplets of nearly equal mass, but differing charge. Thus there are three pions (positive, negative and neutral) and two nucleons, one positive (the proton) and one neutral (the neutron). These are closely analogous to the $2J+1$ spin states of a particle of spin J and may be explained by attributing to each mass multiplet a spin I (isotopic spin) in an abstract space. The different charge states are then interpreted as the different "orientations" in this abstract space and are

specified by the $2I+1$ values of the component I_3 . Thus the pion has isotopic spin 1, with the three charge states specified by $I_3 = \pm 1, 0$. The nucleon has isotopic spin $I = 1/2$, with two charge states $I_3 = \pm 1/2$. Isotopic spin is conserved in strong interactions, just as total angular momentum is conserved. The conservation of angular momentum is related to the invariance of the whole colliding system with respect to its orientation in space. The conservation of isotopic spin is similarly equivalent to the statement that the strong interactions are invariant under a group of two-dimensional unitary transformations, known to group theorists as $SU(2)$.

It has been conjectured that the strong interactions are approximately invariant under a wider group of three-dimensional unitary transformations, $SU(3)$, which combines the notion of isotopic spin with that of hypercharge. According to this theory, the isotopic multiplets of particles of the same spin J can be combined into supermultiplets of roughly similar mass, which form hexagonal or triangular patterns when they are exhibited on a graphical plot of I_3 against Y . By 1963, three such supermultiplets had been established. They form octets of particles, with $J = 0, \frac{1}{2}$ and 1. (These are $\pi^+, \pi^0, \pi^-, \eta^0, K^+, K^0, K^-, \bar{K}^0, \Sigma^+, \Sigma^0, \Sigma^-, \Lambda^0, p, n, \Xi^0, \Xi^-, \rho^+, \rho^0, \rho^-, \phi^0, K^{*+}, K^{*0}, K^{*-}, \bar{K}^{*-}$.) At that time, the known particles of spin $J = 3/2$ formed a tenfold triangular multiplet, with one particle missing—the Ω^- . This missing particle was an isotopic singlet

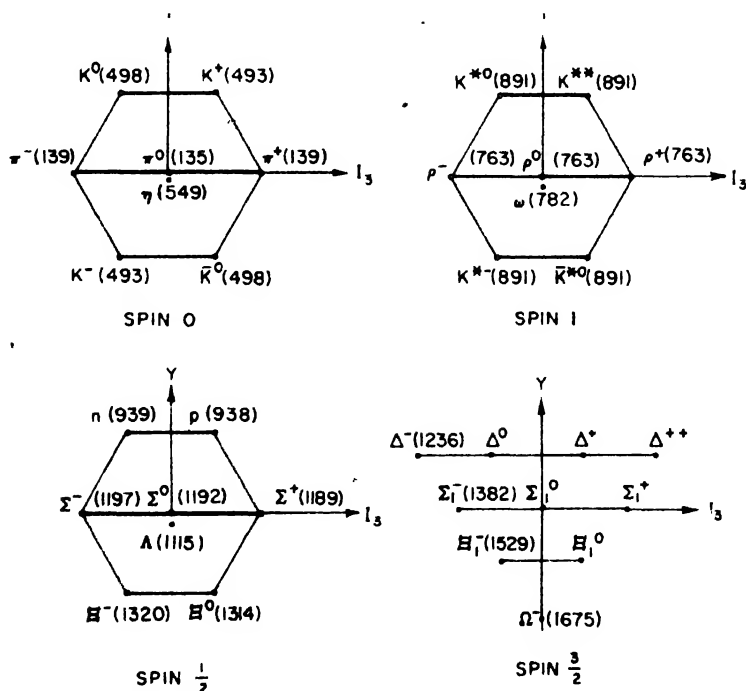


FIG. 1. The well-established $SU(3)$ multiplets. In each of these, particles lying together on horizontal lines form sub-multiplets of $SU(2)$ (isotopic spin). The figures in brackets indicate the masses of the particles in MeV/c^2 .

($I=0$) with hypercharge minus two. It had remarkable physical properties very well defined so that its mass, mode of production, mean life time, and decay could all be predicted. This particle was discovered in 1964 exactly as forecasted, thus confirming the wider unitary scheme.

This scheme provides a kind of "periodic table" for the subnuclear particles. There are clearly very many of them, and the word "elementary" no longer seems appropriate, although it is still in common use. The establishment of "unitary symmetry" for strong interactions has been a great step forward in our understanding of the nuclear world, but it still leaves many questions unanswered. We still have no knowledge of the detailed working of the dynamics of the strong interactions and cannot even tell, for example, why the unitary multiplets, have the masses and spins which they do. There are, however, indications that progress can be made by further application of group theoretic ideas.

There is also the interesting question of whether the subnuclear particles are made up of a few simple entities, just as nuclei are composed of protons and neutrons. If so, unitary symmetry implies that there should be a triplet of particles of nearly equal mass, composed of an isotopic doublet and an isotopic singlet. However in order that the currently known particles should have integer charges (in units of the electron charge), these particles must have fractional charges of $2/3$ and $-1/3$. This is a revolutionary idea and nothing of the kind has been seen, with the energies available with present proton accelerators (about 30 BeV). However, it is conceivable that such particles could be discovered by the construction of yet larger proton accelerators.

P. T. MATTHEWS

Cross-references: ELECTRON, ELEMENTARY PARTICLES, NEUTRINO, NEUTRON, PROTON, WEAK INTERACTIONS.

SUPERCONDUCTIVITY

This article describes the experimental facts relating to superconductivity, our present theoretical understanding, and some practical applications. No attempt is made to follow historical order or to provide an exhaustive treatment; the references at the end of the article will help those who wish to pursue the subject further.

The most spectacular property of a superconductor is the total disappearance of its electrical resistance when it is cooled below a critical temperature T_c . Very careful measurements show that the electrical resistance of a superconductor is at least a factor 10^{17} smaller than the resistance of copper at room temperature and may therefore for all practical purposes be taken to be zero. Some 25 elements and a vast number of alloys and compounds have so far been discovered to be superconducting; examples are In, Sn, V, Mo, Nb-Zr alloys and Nb₃Sn. Transition temperatures range from a few thousandths of

a degree Kelvin (for certain Nb-Mo alloys) all the way to about 18°K for Nb₃Sn.

Another important property is the destruction of superconductivity by the application of a magnetic field equal to or greater than a critical field H_c . This H_c , for a given superconductor, is a function of the temperature given approximately by

$$H_c = H_0(1 - T^2/T_c^2) \quad (1)$$

where H_0 , the critical field at 0°K, is in general different for different superconductors and has values from a few gauss to a couple of thousand gauss. For applied magnetic fields less than H_c , the flux is excluded from the bulk of the superconducting sample, penetrating only to a small depth λ into the surface. The value of λ (called the penetration depth) is in the range 10^{-5} to 10^{-6} cm. Thus the magnetization curve for a superconductor is

$$B \text{ (inside)} = 0 \quad \text{for } H < H_c$$

$$B \text{ (inside)} = B \text{ (outside)} \text{ for } H > H_c$$

This magnetization behavior is reversible and cannot therefore be explained entirely on the basis of the zero resistance. The reversible magnetization behavior is called the Meissner effect.

The existence of the penetration depth λ suggests that a sample having at least one dimension less than λ should have unusual superconducting properties, and such is indeed the case. Thin superconducting films, of thickness d less than λ , have critical fields higher than the bulk critical field, approximately in the ratio of λ to d . This result follows qualitatively from the thermodynamics of the Meissner effect: the metal in the superconducting state has a lower free energy than in the normal state, and the transition to the normal state occurs when the energy needed to keep the flux out becomes equal to this free energy difference. But in the case of a thin film with $d < \lambda$, there is partial penetration of the flux into the film, and thus one must go to a higher applied field before the free energy difference is compensated by the magnetic energy.

It is clear that the existence of the critical field also implies the existence of a critical transport electrical current in a superconducting wire, i.e., that current I_c which produces the critical field H_c at the surface of the wire. For example, in a cylindrical wire of radius r , $I_c = \frac{1}{2}rH_c$. This result is called the Silsbee rule.

All of the above properties distinguish superconductors from "normal" metals. There is another very important distinction, which contains a clue to understanding some of the properties of superconductors. In a normal metal at 0°K, the electrons, which obey Fermi statistics, occupy all available states of energy below a certain maximum energy called the Fermi energy ζ . Raising the temperature of the metal causes electrons to be singly excited to states just above the Fermi energy. There is for all practical purposes a continuum of such excited energy states available above the Fermi energy. The situation is quite different in a superconductor; it turns

out that in a superconductor, the lowest excited state for an electron is separated by an energy gap ϵ from the ground state. The existence of this gap in the excitation spectrum has been confirmed by a wide range of measurements: electronic heat capacity, thermal conductivity, ultrasonic attenuation, far infrared and microwave absorption, and tunneling. The energy gap is a monotonically decreasing function of temperature, having a value $\sim 3.5kT_c$ at 0°K (where k is the Boltzmann constant) and vanishing at T_c (see ENERGY LEVELS).

The superconducting state has a lower entropy than the normal state, and therefore one concludes that superconducting electrons are in a more ordered state. Without, for the present, inquiring more deeply into the nature of this ordering, one can state that a spatial change in this order produced say by a magnetic field will occur, not discontinuously, but over a finite distance ξ , which is called the *coherence length*. The coherence length represents the range of order in the superconducting state and is typically about 10^{-3} cm, though we shall see later that it can in some superconductors take much lower values and lead to some remarkable properties.

Measurements of the transition temperature on different ISOTOPES of the same superconductor showed that T_c is proportional to M^{-1} , where M is the isotopic mass. This isotope effect suggests that the mechanism underlying superconductivity must involve the properties of the lattice, in addition to those of the electrons. Another indication of this is given by the behavior of allotropic modifications of the same element: white tin is superconducting, while grey tin is not, and the hexagonal and face-centered cubic phases of lanthanum have different transition temperatures. A third, and most striking, indication is that the current vs voltage characteristic of a superconducting tunneling junction shows a structure which is intimately related to the phonon spectrum of the superconductor.

The superconducting properties of alloys present a bewildering variety of phenomena. They show a great deal of magnetic hysteresis, with little indication of a perfect Meissner effect. The Silsbee rule is inapplicable, and the resistive transition occurs at fields generally very much higher than in pure superconductors. For example, a wire of Nb_3Sn can carry a current of 10^5 amperes/cm² in an applied field of 100 kilogauss, while a similar wire of lead would carry about 10^3 amperes/cm² in a field of only 100 gauss. When experiments are done using well-annealed (preferably single-crystal) alloys, it is found that the critical currents drop considerably, and the magnetic behavior becomes reversible but still quite unlike that of pure superconductors. The flux is excluded from the interior of the sample up to a well-defined field H_{c1} . When the applied field is raised further, flux begins to penetrate, even though the resistance remains zero, until a second critical field H_{c2} is reached, at which the flux penetration is complete, and normal resistance is abruptly restored.

The theory of superconductivity has developed along two lines, the phenomenological and the microscopic. The phenomenological treatment was initiated by F. London, who modified the Maxwell electromagnetic equations so as to allow for the Meissner effect. His theory explained the existence and order of magnitude of the penetration depth, and gave a qualitative account of some of the electrodynamic properties. The treatment was extended by V. L. Ginzburg and L. D. Landau, and by A. B. Pippard, who in particular emphasized the concept of the range of coherence. A. A. Abrikosov used these ideas to develop a model for alloy superconductors. He showed that if the electronic structure of the superconductor were such that the coherence length ξ becomes smaller than the penetration depth λ , one would get magnetic behavior similar to that observed in alloys, with two critical fields H_{c1} and H_{c2} . The problem of high critical currents in unannealed (or otherwise metallurgically imperfect) alloys and compounds is more complicated because it involves the interaction between the microscopic metallurgical structure and the superconducting properties. This is an area of great research activity because of the technological implication to be mentioned later.

The microscopic theory of superconductivity was initiated by H. Fröhlich, who first recognized the importance of the interactions of electrons with lattice vibrations and in fact predicted the isotope effect before its experimental observation. The detailed microscopic theory was developed by J. Bardeen, L. N. Cooper and J. R. Schrieffer in 1957, and represents one of the outstanding landmarks in the modern theory of solids. The BCS theory, as it is called, considers a system of electrons interacting with the PHONONS, which are the quantized vibrations of the lattice. There is a screened coulomb repulsion between pairs of electrons, but in addition there is also an attraction between them via the electron-phonon interaction. If the net effect of these two interactions is attractive, then the lowest energy state of the electron system has a strong correlation between pairs of electrons with equal and opposite momenta and opposite spin and having energies within the range $k\theta$ (where θ is the Debye temperature) about the Fermi energy. This correlation causes a lowering of the energy of each of these Cooper pairs (named after L. N. Cooper who first pointed out their existence on the basis of some general arguments) by an amount ϵ relative to the Fermi energy. The energy ϵ may be regarded as the binding energy of the pair, and is therefore the minimum energy which must be supplied in order to raise an electron to an excited state. We see thus that the experimentally observed energy gap follows from the theory. The magnitude ϵ_0 of the gap at 0°K is

$$\epsilon_0 \approx 4k\theta \exp \left(-\frac{1}{NV} \right)$$

where N is the density of electronic states at the Fermi energy and V is the net electron-electron

interaction energy. The superconducting transition temperature T_c is given by

$$3.5kT_c \approx \epsilon_0$$

It has been shown that the BCS theory does lead to the phenomenological equations of London, Pippard and Ginzburg and Landau, and one may therefore state that the basic phenomena of superconductivity are now understood from a microscopic point of view, i.e., in terms of the atomic and electronic structure of solids. It is true, however, that we cannot yet, *ab initio*, calculate V for a given metal and therefore predict whether it will be superconducting or not. The difficulty here is our ignorance of the exact wave functions to be used in describing the electrons and phonons in a specific metal, and their interactions. However, we believe that the problem is soluble in principle at least.

The range of coherence follows naturally from the BCS theory, and we see now why it becomes short in alloys. The electron mean free path is much shorter in an alloy than in a pure metal, and electron scattering tends to break up the correlated pairs, so that for very short mean free paths one would expect the coherence length to become comparable to the mean free path. Then the ratio $\kappa \approx \lambda/\xi$ (called the Ginzburg-Landau order parameter) becomes greater than unity, and the observed magnetic properties of alloy superconductors can be derived. The two kinds of superconductors, namely those with $\kappa < 1/\sqrt{2}$ and those with $\kappa > 1/\sqrt{2}$ (the inequalities follow from the detailed theory) are called respectively type I and type II superconductors.

There have been several attempts at technological applications of superconductors. The most spectacularly successful one is the use of certain type II superconductors like Nb-Zr and Nb-Ti alloys, and Nb₃Sn, in making electromagnets. In a conventional electromagnet employing normal conductors, the entire electric power applied to the magnet is consumed as Joule heating. For a magnet to produce 100 kilogauss in a reasonable volume, the power requirement can run into megawatts. In striking contrast, a superconducting magnet develops no Joule heat because its resistance is zero. Indeed, if such a magnet has a superconducting shunt placed across it after it is energized, the external power supply can be removed, and the current continues to flow indefinitely through the magnet and shunt, maintaining the field constant. Superconducting magnets have already been constructed producing fields of over 100 kilogauss in usable volumes. There is a natural upper limit to the critical field possible in such superconductors, given by the paramagnetic energy of the electrons (due to their spin moment) in the normal state becoming equal to the condensation energy of the Cooper pairs in the superconducting state. This leads to a limit of about 360 kilogauss for a superconductor with a T_c of 20° K.

The possibility of a persistent current in a superconducting ring, and the sharpness and

speed of the change in resistance at the superconducting to normal transition, have led to consideration of superconductors for memory and logic circuits in computers. Further work, mainly of a technological nature, is needed before practical superconducting computers can be built.

B. S. CHANDRASEKHAR

References

- Lynton, Ernest A., "Superconductivity," New York, John Wiley, 1962. A concise, extremely readable book in which most of the ideas in this article are elaborated.
- Shoenberg, D., "Superconductivity," New York, Cambridge University Press, 1952. An excellent account of the field up to about 1952.
- London, F., "Superfluids," Vol. I, New York, John Wiley, 1950. A classic account of the early theory of superconductivity.
- Schrieffer, J. Robert, "Theory of Superconductivity," New York, W. A. Benjamin, 1964. An advanced and comprehensive account of the microscopic theory.
- Rev. Mod. Phys.* **36**, 1-331 (1964). This contains the proceedings of an international conference on superconductivity held in 1963.
- Advan. Cryog. Eng.* **7** (1962); **8** (1963); **9** (1964). This is a continuing series containing the proceedings of the annual Cryogenic Engineering Conference, and deals with the many applications of superconductivity, as well as the techniques of low temperatures.

Cross-references: CONDUCTIVITY, ELECTRICAL; ELECTRON SPIN; ENERGY LEVELS; FERMI-DIRAC STATISTICS AND FERMIONS; HEAT CAPACITY; HEAT TRANSFER; MAGNETISM; PHONONS; SEMICONDUCTORS; SOLID-STATE PHYSICS; SUPERFLUIDITY.

SUPERFLUIDITY

Superfluidity is a term used to describe a property of condensed matter in which a resistanceless flow of current occurs. The mass-four isotope of helium in the liquid state plus some twenty-three metallic elements are presently known to exhibit this phenomenon. In the case of liquid helium, these currents are hydrodynamic; for the metallic elements, they consist of electron streams. The effect occurs only at very low temperatures in the vicinity of the absolute zero (-273.16°C or 0°K). In the case of helium, the maximum temperature at which the effect occurs is about 2.2°K ; for metals the highest temperature is in the vicinity of 20°K .

If one of these metals (called superconductors) is cast in the form of a ring and an external magnetic field is applied perpendicular to its plane and then removed, a current will flow round the ring induced by Faraday induction. This current will produce a magnetic field, proportional to the current, and the size of the current may be observed by measuring this field. Were the ring (e.g.,

one made of Pb) at a temperature above 7.2°K, this current and field would decay to zero in a fraction of a second. But with the metal at a temperature below 7.2°K before the external field is removed, this current shows no signs of decay even when observations extend over a period of a year. As a result of such measurements, it has been estimated that it would require 10^{99} years for the supercurrent to decay! To the best of our knowledge, therefore, the lifetime of these "persistent" currents is infinite. The persistent or frictionless currents in superconductors are not a recent discovery, they were observed first some 50 years ago (see SUPERCONDUCTIVITY).

In the case of liquid helium, these currents are, as mentioned, hydrodynamic, i.e., they consist of streams of neutral (uncharged) helium atoms flowing in rings. Since, unlike electrons, the helium atoms carry no charge, there is no resulting magnetic field. This makes such currents much more difficult to create and detect. Nevertheless, as a result of research carried out here and in England in the past five or six years, the existence of these supercurrents in liquid helium has definitely been proved. These currents have very recently (1964) been observed for periods as long as 12 hours, a time of the order of 10^3 shorter than is the case for electron currents. Nevertheless, our present belief is that these hydrodynamic currents also possess an infinite lifetime.

As mentioned, the empirical discovery of infinite-lifetime electron currents is of considerable antiquity, and from the beginning, many attempts have been made to explain the effect theoretically. Until very recently all such attempts have failed completely, and as a matter of fact, the theoretical picture is still not completely satisfactory. Nevertheless, immense progress in this direction has occurred in the past decade largely as the result of work in the U.S.A., England and Russia.

Although superfluidity in liquid helium is important to our basic understanding of the phenomenon, it is the effect in superconductors which arouses the most interest. This is due, in part at least, to the possible practical applications to which the effect might lead. In ordinary conductors (e.g., copper), the flow of an electric current is always accompanied by energy dissipation. The supercurrents propagate with no such power loss; superconducting transmission lines would be an economic advance of the first order. However, they would require "room temperature" superconductors, and unfortunately, modern theories, while they do not absolutely prohibit such, render their occurrence most unlikely.

In other ways, however, superconductors have already proved of practical value. Since the currents once created are there for all time, they have fairly obvious use as memory elements in computers. Again the persistent currents form a sort of super gyroscope more perfect than any so far devised. In recent years alloy superconductors have been found (e.g., Nb₃Sn) which can support persistent currents even in the presence of intense

magnetic fields of the order of 100 000 gauss. It is, generally speaking, a property of elemental superconductors that the persistent current is quenched in fields of a few hundred gauss; the situation in some alloys like the above is very different.

This has led to the development of intense field solenoidal magnets which maintain their magnetic fields with zero energy dissipation. A conventional electromagnet of the same size would consume many hundreds of kilowatts. This discovery has important consequences in several areas of physics.

Both aspects of superfluidity find their theoretical explanation in what is called a two-fluid theoretical model. We suppose that liquid helium and superconductors are quantum systems possessing a zero energy ground state plus a series of available states of higher energy called normal or excited states. The occupancy of the ground state is zero above the superfluid transition temperature (2.2 K for liquid helium) but grows steadily as the temperature is lowered. At absolute zero, the whole system is in the ground state. At any finite temperature, below transition, the system is in a mixture of ground and excited states. It is the particles in the ground state which form the persistent current.

A simple, though not entirely accurate, *raison d'être* for the persistent currents lies in the fact that in the ground state, all the very many particles possess the same single wave function. It follows from this that such particles are not easily scattered out of the ground state—a finite amount of energy is required. Such an assembly will not readily interact with outside particles including those in excited states. Since fluid friction or viscosity is due to particle momentum interchange between neighboring layers of liquid, it follows that particles in the ground state will possess zero viscosity. This, in turn, means frictionless flow.

An assembly of particles obeying Bose-Einstein statistics will, below a certain temperature, possess a ground state like the one postulated above. It is known that Bose-Einstein statistics apply only to particles which possess zero (or integral) spin. The neutral He⁴ atom possess zero spin. The isotope He³ (spin 1/2) does not; and, in fact, liquid He³ is not a superfluid. But this is also true of electrons which possess half integral spin and obey a very different statistic (Fermi-Dirac). In this statistic no more than two electrons can possess the same wave function. Hence a ground state, in the above sense, can clearly not exist. Thus according to our model, superfluid flow of electrons cannot occur—but it does!

A way out of this dilemma was suggested nearly 20 years ago. Namely, combine the electrons into pairs, with opposite spins. The resulting "particle" is then a boson and may properly reside in the ground state. The reason why this idea was not accepted by theoretical physicists until very recently was because no mechanism could be found by which the electrons could be induced to form pairs. It turns out that the required mechanism arises as a result of a phonon

(quantized lattice wave) emitted by one electron and absorbed by another at some other place. This couples the two together.

As mentioned, none of the current theories clearly explains persistent currents. It is not at all evident that their lifetimes should be infinite. It is thought by some that a better formulation for statistical mechanics than presently exists may be necessary before this becomes possible.

A system of persistent currents distinguished by the fact that many particles possess the same wave function constitutes a quantum effect on a hitherto unknown macroscopic scale. In other words, superfluids should exhibit quantum effects of such size that they are readily amenable to experimentation. Within the past few years, two such effects, one in each of the superfluids, have been found. A long thin cylinder with a hole along the axis (i.e., a tube) is similar to the previously mentioned superconducting ring in that persistent currents may also be produced in it. If the length of the cylinder is large compared to the diameter of the hole, the "trapped" magnetic field due to the persistent current is substantially uniform. The flux is the product of this field by the area of the hole.

As a consequence of the fact that the particles producing the current all possess the same wave function, it may be shown that this flux is quantized in integral multiples of hc/q (h = Planck's constant, c = velocity of light, q = charge on the particles).

This prediction has recently (1961) been confirmed experimentally. Further, these experiments show that $q = 2e$, where e is the charge on an electron. Thus the pair hypothesis is very nicely proven.

An interesting consequence of the above effect is that it is impossible to have any field (below a certain value depending on the size of the hole) at all in the cavity. Thus the measurements show that $hc/q \approx 2 \times 10^{-7}$ so that with a hole 1μ in diameter, a field less than about 13 gauss could not exist. We should therefore have a truly field-free space.

A very similar situation, due fundamentally to the same cause, exists for the hydrodynamic currents in liquid helium. In this case, it is the hydrodynamic "circulation" which is quantized in units of h/m (h = Planck's constant, m = helium atom mass). By circulation we mean $\oint \mathbf{v} \cdot d\mathbf{l}$ where \mathbf{v} is the velocity of the atom in the ring-current and $d\mathbf{l}$ is a line element of the periphery of the ring. This effect has also been observed quite recently (1958) in laboratory experiments.

There is also a surprising consequence connected with this effect. Suppose the helium was being rotated in its containing vessel at a temperature above 2.2°K. Here the helium is "classical" and would rotate, like any other familiar liquid, at the same angular velocity as the vessel. Suppose, now, that the helium was cooled down to near 0°K with the vessel still rotating. If the initial speed were less than that required to produce one quantum of circulation, the rotating

helium would come to rest while the container continued to rotate!

This is analogous to the behavior of the magnetic flux on the superconductor, the angular velocity being the quantity analogous to the magnetic field. To be sure, the above experiment has, to date, not been performed, but this is entirely due to difficulties with the existing instrumentation. With advances in technique, it seems very likely that the experiment can eventually be performed, and there is little doubt that the result predicted will be observed.

C. T. LANE

Cross-references: BOSE-LINSTEIN STATISTICS AND BOSONS; CONDUCTIVITY, ELECTRICAL; ELECTRON SPIN; FERMI-DIRAC STATISTICS AND FERMIONS; PHONONS; SUPERCONDUCTIVITY.

SURFACE PHYSICS

Thermodynamics and Simplified Models of Surfaces. The reversible work required to create a unit area of solid surface at a constant volume, temperature and chemical potential is customarily called the surface tension and denoted as γ . The Helmholtz free energy per unit area of surface, f_s , is thus:

$$f_s = \gamma + \sum_i \mu_i \Gamma_i$$

where μ_i and Γ_i are respectively the chemical potential and excess surface density of the i th component. For one-component systems, $\gamma = f_s$, Γ_i then being 0. For a liquid, γ is equal to the surface stress; in a solid, this is not true, and in fact, components of the surface stress tensor may be zero or of either sign. The above definitions are somewhat inadequate for a general surface. The ambiguities in the definition of surface properties have been treated by Gibbs and discussed in several places.^{1,2}

Crudely the surface tension arises from the fact that the surface atoms are bound to fewer neighbors than are atoms in the bulk phase. Assuming neighbor-neighbor interactions and no surface distortion, one may estimate γ by counting broken bonds which yields, for a close packed surface, $A\gamma \approx 1/4 \frac{L_0}{N}$ where A is the area per surface atom, L_0 is molar heat of sublimation and N is Avagadro's number. These assumptions are obviously inadequate, and the estimate gives only the order of magnitude of γ which for metals is on the order of several hundred to several thousand ergs per square centimeter.

For a crystal, γ is a function of the crystallographic orientation of the surface. A surface with a mean orientation the same as that of a low index face is usually thought to be smooth. For surfaces inclined with respect to low index planes there will be atomic planes which terminate on the surface at steps. In the crude picture, additional broken bonds are associated with the steps and

thus the energy will be higher than for a close-packed plane. Again, counting broken bonds, and assuming steps widely separated so as not to interact, leads to:

$$\gamma(\theta) = \gamma_0 \cos \theta + \alpha \sin |\theta|; \theta \ll 1$$

where γ_0 is the surface tension of the low index face, α is proportional to the energy per step, and $\gamma(\theta)$ is the surface tension for a surface inclined by an angle θ from the low index face. Thus, on a polar plot of $\gamma(\theta)$ vs the orientation of the normal to the surface, there will be cusps at the orientations of close-packed planes. At finite temperature, the cusps disappear because of entropy considerations.

The surface tension is changed by adsorption on the surface in accordance with Gibbs' adsorption equation:

$$d\gamma = -s_s dT - \sum_i \Gamma_i d\mu_i$$

where s_s is the specific surface entropy. Thus, at constant temperature, introduction of a component which is adsorbed on the surface reduces γ and generally changes the shape of the γ plot since the Γ_i depend on the orientation of the surface.

The equilibrium shape of a crystal depends on the orientation dependence of γ and may be determined graphically by the Wulff construction. In the polar plot of the γ diagram mentioned above one constructs planes perpendicular to all radius vectors at the point at which they intersect the γ plot. Then the figure containing all points which may be reached from the origin without crossing any such plane has the same shape as the equilibrium crystal. Herring¹ has discussed proofs of the Wulff construction and possible equilibrium shapes of crystals.

A smooth surface of an orientation not represented in the Wulff equilibrium shape will be able to reduce its surface energy by developing facets, keeping the same mean orientation. Such faceting may be observed if the kinetics allow it in reasonable time and if other processes such as selective evaporation or chemical reactions do not allow the development of the equilibrium surface. When faceting driven by the surface energy is observed, it should yield information about the anisotropy of γ .

Atoms may be transported from one surface site to another by (1) transport through a vapor phase, (2) diffusion through the bulk, and (3) surface diffusion. Only the last is considered briefly here. The surface self-diffusion is thermally activated and often is much faster than bulk diffusion and vapor transport. The situation is somewhat complicated by the fact that atoms may occupy a variety of sites on a surface, i.e., (1) as an adatom on an otherwise smooth surface, (2) as an atom on a step, (3) as an atom at a jog in the step, (4) as an atom in a step, or (5) as an atom within a smooth layer. A diffusing atom will occupy each type of these sites in its motion over the surface, and for each type, there will be a

different average jump frequency. In a given experimental situation, the diffusion will depend on the population of atoms in the various sites which may not be in equilibrium. Furthermore, the diffusion will not necessarily be isotropic in the plane of the surface and will be affected by adsorbed impurities.

The surface self diffusion is measured experimentally by observing development of grain boundary grooves, smoothing of scratches, thermal faceting, blunting of field-emission microscope tips, by radioactive tracer techniques and by field ion-microscopy. N. A. Gjostein³ has written a review of surface diffusion and has summarized available experimental data.

Atomic Structure of Surfaces. In the previous paragraphs, reference has been made to surface structures as if they resulted by simply terminating the bulk structure along a plane with perhaps some terraces, steps and jogs. Recently, mainly through slow electron diffraction experiments, considerable detailed information concerning actual structures has become available. Within several atomic layers of the surfaces, real structures may differ from those of the bulk by small distortions or may be completely "reconstructed." Elizabeth A. Wood has proposed conventions for describing surface structures which are followed here.⁴ The region of the material in which the structure is periodic in the direction normal to the surface (the structure is triperiodic) is the substrate. The region in which the structure is diperiodic is called the selvage. Whereas in a triperiodic structure unit cells are arranged in one of the fourteen Bravais lattices, in a diperiodic structure unit meshes are arranged in one of five nets. The five nets are described in Table I.

Most of the usual crystallographic notation of triperiodic structures can be carried over to diperiodic systems with only obvious modifications. These are listed by Wood. The reciprocal lattice becomes a family of rods normal to the net. The Miller indices of rows are customarily given in terms of the substrate, and since surface structures often have large unit meshes, fractional Miller indices are used.

The surface structure usually bears a close relationship to the substrate, and this fact is used in a convenient notation to designate the surface structure. The orientation of the substrate plane parallel to the surface is specified first, then the ratios of the lengths of unit mesh vectors of the surface structures and of the substrate plane, and finally the chemical symbol of any adsorbed atoms if present. As an example, Ni(110) 3×1 O refers to a surface structure on the (110) face of nickel with unit mesh vectors parallel to the substrate, with

$$|\vec{a}_{\text{surface}}|/|\vec{a}_{\text{substrate}}| = 3,$$

$$\text{and } |\vec{b}_{\text{surface}}|/|\vec{b}_{\text{substrate}}| = 1$$

and containing adsorbed oxygen. If the two ratios above are the same, only their common value is given.

TABLE I (TAKEN FROM WOOD'S PAPER)*

Shape of mesh	Lattice symbol†	Choice of Axis	Nature of Axis and Angles	Name of Corresponding System
General parallelogram	p	$a \neq b$	$a \neq b$ $\gamma \neq 90^\circ$	Oblique
Rectangular	p	Shortest two mutually perpendicular vectors $a \neq b$	$a \neq b$ $\gamma \neq 90^\circ$	Rectangular
Square	p	Shortest two mutually perpendicular vectors	$a = b$ $\gamma = 90^\circ$	Square
120° angle rhombus	p	Shortest two vectors at 120° to each other	$a = b$ $\gamma = 120^\circ$	Hexagonal

* From Wood, Elizabeth A., "Vocabulary of Surface Crystallography," *J. Appl. Phys.*, 35, 1306 (1964).

† p refers to a primitive mesh and c to a centered one.

The surface structure of several materials and adsorbed layers on them have been investigated by low-energy electron diffraction techniques. Summaries of the results and references to them are given by J. J. Lander⁵ and in reference 6.

Surfaces of the dense planes of metals, layered materials such as graphite and cleavage planes of ionic crystals generally differ by only small distortions from the corresponding bulk structures. As examples: Peria and Johnson⁶ find that the low-energy electron diffraction from the cleavage plane of MgO is best fit by displacing the second layer of magnesium ions 0.35Å toward the surface, the oxygen ions remaining fixed; Germer and Mac Rae⁸ suggest that the spacing between the first and second atomic layers of clean nickel is 5 per cent greater than in the bulk material. These results should perhaps be considered as tentative since a completely satisfactory understanding of the scattered intensities in slow electron diffraction experiments is not yet available.

Covalently bonded materials are more likely to have "reconstructed" surface structures as might be expected from the unfavorable energy of structures with broken bonds. There may be a wide variety of such structures for surfaces of each orientation, and prediction of such structures from first principles is as yet impossible. Interesting examples of such structures are the surfaces of freshly cleaved and then annealed silicon. In a (111) surface cleaved at room temperature in vacuum, alternate "rows of top layer atoms are brought into the adjacent rows and given double bonds to the atoms already there. The broken bonds of atoms in the row below are then restored by displacing these atoms to form paired rows."⁹ It is interesting that this structure has twofold rotational symmetry on a substrate of threefold symmetry. At 700°C, the structure is "reconstructed" to a more stable one, Si (111)-7 or Si (111)-5 which are made up of "warped benzene rings", each unit having three atoms from both the first and second layers of atoms. The outer layer contains only about three-fourths as many atoms as the freshly cleaved structure, and

therefore diffusion of atoms out of the surface is required for its formation.

The variety and complexity of actual surface structures illustrates the serious shortcoming of those calculations of surface properties based on simplified models.

Electronic Structure. The electronic structure in the vicinity of the surface differs from that which would result from simply terminating the crystal along a plane and allowing no subsequent relaxation. In a metal, the electronic charge distribution does not end abruptly at the surface, but decreases smoothly and extends beyond the boundaries of the Wigner-Seitz cells of the surface atoms. This leads to a surface dipole layer, negative on the outside, which, along with the image potential, is the major contribution of the surface to the work function of the metal. This contribution is the order of a few tenths of an electron volt and is the order of 10 per cent of the observed work function. The larger contributions are due to the bulk properties and are not discussed here. In a metal, the thickness of the region in which the electron density differs from its bulk value is on the order of the interatomic spacing, which is of course the order of the wavelength of the most energetic conduction electrons. This smooth decrease of the charge density at the surface is a result of a compromise between the higher electrostatic energy of the dipole layer and the kinetic energy it would require to make a steeper boundary. This same compromise will make the boundary of the electron distribution smoother in the directions parallel to the surface than the boundary of the ion cores. The hill and valley structure of atomically rough surfaces then gives an additional dipole layer with the positive side outward. Thus, rougher surfaces have lower work functions than more closely packed, smoother surfaces.

The electronic distribution makes a contribution to the surface energy, but realistic calculations are very difficult and the usual discussions are for a free electron gas model. Because the conduction electrons adjust easily to the potential at the surface, metals have surface energies of only 1/2 to 2/3 that which

would be predicted from the sublimation energy. For both the surface energy and the work function, the lattice potential and correlation and exchange terms are important. The theoretical approaches to these problems are reviewed by Ewald and Juretschke¹⁰ and by Herring.¹¹

For the half-infinite crystal with a periodic potential, there is the possibility that there are localized states on the surface with energies within the band gaps of the bulk structure. The wave functions of these states are exponentially damped on both sides of the surface. The area density of such states should be the same order as the density of surface atoms. They are called Tamm states after I. Tamm who first investigated them for a one dimensional Kronig-Penney model. The conditions for the existence of surface states for actual crystals are not completely clear, but in some models they appear if the energy bands of the bulk overlap or if the surface atomic structure is different from the bulk, resulting in a potential trough. Recent discussions of both theoretical and experimental aspects of surface states can be found in the papers of reference 6.

Surface states are particularly important in semiconductors, and there has been extensive work on Germanium crystals¹². They are of two types: (1) "fast" states which occur on atomically clean surfaces or at the interface between the semiconductor and its oxide layer, which are thus in good electrical contact with the bulk, and which therefore may adjust their occupation in the order of 10^{-6} to 10^{-8} second following a disturbance; and (2) "slow" states on the outer surface of the oxide layer and for which changes in the occupation after a disturbance are much slower. The observed density of the "fast" states is a few orders of magnitude smaller than the surface density of atoms so they are not truly Tamm states, but may be due to imperfections or deformations of the surface.

There will be a surface space charge region which shields any charge localized in the surface states and which will, because of the low density of free carriers in the semiconductor, extend into the bulk to depths usually the order of 10^{-5} cm. The energy bands bend through this region in order to give the requisite shielding charge. The bending of the bands will lead to regions in which the conductivity type (n or p) may be opposite that of the bulk (i.e., an inversion layer) or in which the conductivity is more strongly of the same type as the bulk (i.e., an accumulation layer) or, finally, in an intermediate case, in a region where the conductance of the specimen is decreased, called an exhaustion layer. The energies, densities, and trapping and recombination cross sections of the surface states can be investigated by measurements of various electrical properties. The discussion of these investigations is beyond the scope of this article and the reader is referred to recent reviews by Watkins,¹² Law,¹³ and by Zemel.¹⁴

References

1. Herring, C., in Gomer and Smith, Eds., "Structure and Properties of Solid Surfaces," Chicago, Ill., University of Chicago Press, 1952.
2. Mullens, W. W., in "Metal Surfaces," ASM, 1962.
3. Gjostein, N. A., in "Metal Surfaces," ASM, 1962.
4. Wood, E. A., *J. Appl. Phys.*, **35**, 1306 (1964).
5. Lander, J. J., *Surface Sci.*, **1**, 125 (1964).
6. *Surface Science*, **2** (1964). This entire volume contains the proceedings of the International Conference on the Physics and Chemistry of Solids (1964).
7. Johnson, D. C., Ph.D. thesis, University of Minnesota, 1964 (unpublished).
8. Mac Rae, A. U., and Germer, L. H., *Ann. N.Y. Acad. Sci.*, **101**, 627 (1963).
9. Lander, J. J., Gobeli, G. W., Morrison, J., *J. Appl. Phys.*, **34**, 2298 (1963).
10. Ewald, P. P. and Juretschke, H., in Gomer and Smith, Eds., "Structure and Properties of Solid Surfaces," Chicago, Ill., University of Chicago Press, 1952.
11. Herring, C., in "Metal Interfaces," A.S.M., 1952.
12. Watkins, T. B., "Progress in Semiconductors," New York, John Wiley & Sons, Vol. 5, 1960.
13. Law, J. T., in Hannay, Ed., "Semiconductors," New York, Reinhold Publishing Corp., 1959.
14. Zemel, J. N., *Ann. N.Y. Acad. Sci.*, **101**, 830 (1963).

Cross-references: CRYSTALLOGRAPHY, SOLID-STATE PHYSICS, SOLID-STATE THEORY, SURFACE TENSION.

SURFACE TENSION

Surface tension results from the tendency of a liquid surface to contract. It is given by the tension σ across a unit length of a line on the surface of a liquid. The surface tension of a liquid depends on the temperature and diminishes as temperature increases and becomes 0 at the critical temperature. For water σ is 0.073 newtons/m at 20°C, and for mercury it is 0.47 newtons/m at 18°C.

Surface tension is intimately connected with capillarity, that is, rise or depression of liquid inside a tube of small bore when the tube is dipped into the liquid. Another factor which is related to this phenomenon is the angle of contact. If a liquid is in contact with a solid and with air along a line, the angle θ between the solid-liquid interface and the liquid-air interface is called the angle of contact (Fig. 1). If $\theta = 0$, the liquid is said to wet the tube thoroughly. If θ is less than 90°, the liquid rises in the capillary, and if θ is greater than 90°, the liquid does not wet the solid but is depressed in the tube. For mercury on glass, the angle of contact is 140°, so that mercury is depressed when a glass capillary is dipped into mercury. The rise h of the liquid in the capillary is given by $h = 2\sigma \cos \theta / r \rho g$, where r is the radius of the tube, ρ the density of the liquid, and g is the acceleration due to gravity.

Surface tension can be explained on the basis of molecular theory. If the surface area of liquid is

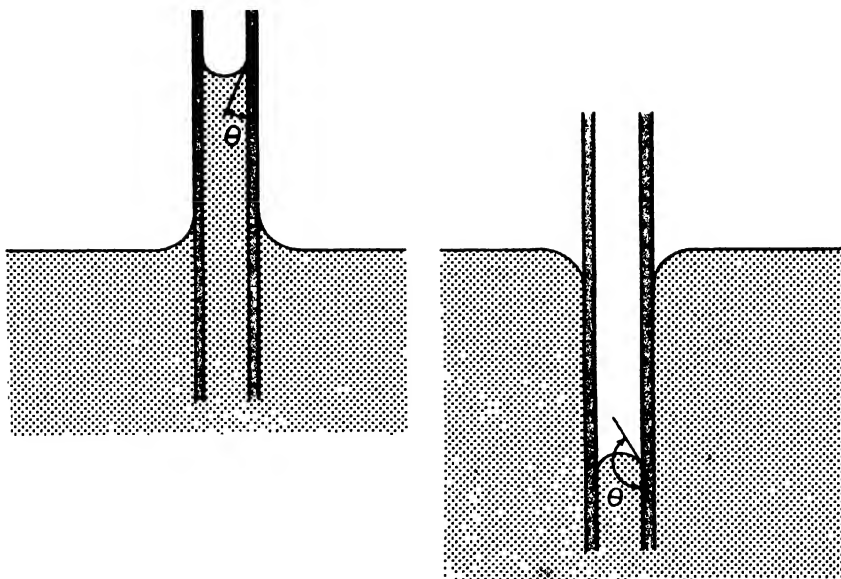


FIG. 1. Capillarity.

expanded, some of the molecules inside the liquid rise to the surface. Because a molecule inside a mass of liquid is under the forces of the surrounding molecules while a molecule on the surface is only partly surrounded by other molecules, work is necessary to bring molecules from the inside to the surface. This indicates that force must be applied along the surface in order to increase the area of the surface. This force appears as tension on the surface and when expressed as tension per unit length of a line lying on the surface, it is called the surface tension of the liquid.

The molecular theory of surface tension has been dealt with since the time of Laplace (1749–1827). As a result of the clarification of the nature of intermolecular forces by quantum mechanics and of the recent development in the study of molecular distribution in liquids, the nature and the value of surface tension have come to be understood from a molecular point of view.

Surface tension is closely associated with a sudden but continuous change in the density from the value for bulk liquid to the value for the gaseous state in traversing the surface (Fig. 2). As a result of this inhomogeneity, the stress across a strip parallel to the boundary— p_N per unit area—is different from that across a strip perpendicular to the boundary— p_T per unit area. This is in contrast with the case of homogeneous fluid in which the stress across any elementary plane has the same value regardless of the direction of the plane.

The stress p_T is a function of the coordinate z , the z -axis being taken normal to the surface and directed from liquid to vapor. The stress p_N is constant throughout the liquid and the vapor. The figure shows the stress p_N and p_T . The stress

$p_T(z)$ as function of z is also shown on the left side of the Figure.

The surface tension is given by integrating the difference $p_N - p_T(z)$ over z :

$$\sigma = \int_{(\text{liquid})}^{(\text{vapor})} [p_N - p_T(z)] dz$$

A statistical mechanical treatment of the system leads to the expression of p_N and p_T in terms of intermolecular forces, density distribution, and

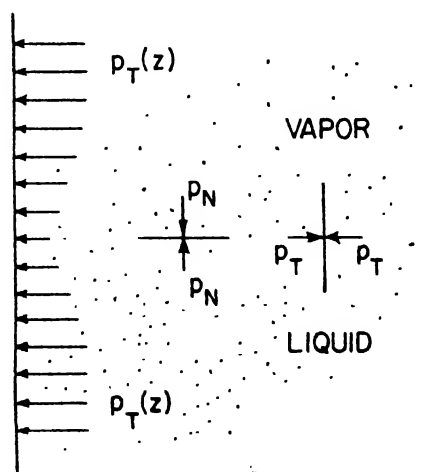


FIG. 2. Stress relationships in surface tension.

the distribution of other molecules around a molecule which is located at a position z . The change of the number density and the distribution

of other molecules around a central molecule at z are problems which have not yet been completely solved. Some simplifying assumptions such as to assume the transition layer to be a mathematical surface of density discontinuity have made the theory more amenable to numerical calculations. It can be said that so far as such simple liquids as liquid argon are concerned, the values of surface tension have been calculated theoretically in fair agreement with observed values.

AKIRA HARASIMA

References

- Hirschfelder, J. O., Curtiss C. F., and Bird, R. B., "Molecular Theory of Gases and Liquids," p. 336, New York. John Wiley & Sons, 1954.
 Harasima, A., "Molecular Theory of Surface Tension," in Prigogine, I., Ed., *Advan. Chem. Phys.*, 1, (1958).
 Ono, S., and Kondo, S., "Molecular Theory of Surface Tension in Liquids," in Flügge, S., Ed., "Encyclopedia of Physics," Vol. X, "Structure of Liquids," p. 134, Berlin, Springer.

Cross-references: LIQUID STATE, STATISTICAL MECHANICS.

SYMBOLS, UNITS AND NOMENCLATURE IN PHYSICS

Introduction. International communication and cooperation in science continue to grow in importance. Not the least important aspect of this cooperation is uniformity of international usage of symbols, units, and nomenclature in physics. The proliferation of research in countries throughout the world makes the problem of uniformity of usage a vital one in the dissemination of scientific literature, for it is obvious that much time and effort can be wasted in misunderstandings arising from terminology.

The recommendations given here are primarily those of the Commission on Symbols, Units and Nomenclature of the International Union of Pure and Applied Physics* as approved by the General Assemblies of IUPAP in 1960 and 1963. There is agreement on most of these recommendations among the following international organizations:

- (1) International Organization for Standardization, Technical Committee 12;
- (2) General Conference on Weights and Measures;
- (3) International Union of Pure and Applied Chemistry;
- (4) International Electrotechnical Commission, Technical Committees 24, 25;
- (5) International Commission on Illumination.

Physical Quantities—General Recommendations. A physical quantity, represented by a symbol, is equivalent to the product of a numerical value and a unit. For dimensionless physical quantities, the unit often has no name or symbol and is not explicitly indicated.

* UIP 9; SUN 61-44; reproduced in *Phys. Today*, 15, No. 6, 19 (June 1962).

EXAMPLES:

$$E = 200 \text{ erg} \quad n_{\text{qu.}} = 1.55$$

$$F = 27 \text{ N} \quad \nu = 3 \times 10^8 \text{ Hz}$$

Symbols for Physical Quantities—General Rules.

(1) Symbols for physical quantities should be single letters of the Latin or Greek alphabet with or without modifying signs: subscripts, superscripts, dashes, etc.

REMARK: (a) An exception to this rule consists of the two-letter symbols, which are sometimes used to represent dimensionless combinations of physical quantities. If such a symbol, composed of two letters, appears as a factor in a product, it is recommended to separate this symbol from the other symbols by a dot or by brackets or by a space.

(b) Abbreviations, i.e., shortened forms of names or expressions, such as p.f. for partition function should not be used in physical equations. These abbreviations in the text should be written in ordinary Roman type.

(2) Symbols for physical quantities should be printed in *italic* (i.e., sloping) type.

REMARK: Subscripts and superscripts should be in italic type when they are symbols for physical quantities or when they are running indices: e.g., C_p where p represents pressure but C_k where g means gas; g_{ik} where i and k are running indices but E_k where k means kinetic.

(3) Symbols for vectors and tensors: Special type fonts are recommended for these quantities but not for their components:

(a) Vectors should be printed in boldface type.

(b) Tensors of the second rank[†] should be printed in sans serif type.

* *Simple Mathematical Operations.* (1) Addition and subtraction of two physical quantities are indicated by:

$$a + b \text{ and } a - b$$

(2) Multiplication of two physical quantities may be indicated in one of the following ways:

$$ab \quad a b \quad a \cdot b \quad a \times b$$

REMARK: The various products of vectors and tensors may be written in the following ways:

Scalar product of vectors A and B : $A \cdot B$

Vector product of vectors A and B : $A \times B$

Dyadic product of vectors A and B : AB

Scalar product of tensors S and T ($\sum_k S_{ik} T_{ki}$): $S : T$

Tensor product of tensors S and T ($\sum_k S_{ik} T_{ki}$): $S \cdot T$

Product of tensor S and vector A ($\sum_k S_{ik} A_k$): $S \cdot A$

(3) Division of one quantity by another quantity may be indicated in one of the following ways:

$$\frac{a}{b} \quad a/b \quad a b^{-1} \quad a(1/b)$$

These procedures can be extended to cases where one of the quantities or both are themselves products, quotients, sums or differences of other quantities.

If necessary, brackets have to be used in accordance with the rules of mathematics.

If the solidus is used to separate the numerator from the denominator and if there is any doubt where the numerator starts or where the denominator ends, brackets should be used.

EXAMPLES:

Expressions with a Horizontal Bar Same Expressions with a Solidus

$$\frac{a}{bcd}$$

$$a/bcd$$

$$\frac{2}{9} \sin kx, \frac{1}{2} RT$$

$$(2/9) \sin kx, (1/2)RT \text{ or } RT/2$$

$$\frac{a}{b} \cdots c$$

$$a/b \cdots c$$

$$\frac{a}{b-c}$$

$$a/(b-c)$$

$$\frac{a-b}{c-d}$$

$$(a-b)/(c-d)$$

$$\frac{a}{c} \cdots \frac{b}{d}$$

$$a/c \cdots b/d$$

REMARK: It is recommended that in expressions like:

$$\sin \{2\pi(x-x_0)/\lambda\}, \quad \exp \{(r-r_0)/\sigma\},$$

$$\exp \{-V(r)/kT\}, \quad \sqrt{\epsilon/c^2}$$

the argument should always be placed between brackets, except when the argument is a simple product of two quantities, e.g., $\sin kx$. When the horizontal bar above the square root is used no brackets are needed.

Units—General Recommendations. *Symbols for Units—General Rules.* (1) *Symbols for units of physical quantities should be printed in roman (upright) type.*

(2) *Symbols for units should not contain a final full stop and should remain unaltered in the plural, e.g., 7 cm and not 7 cms.*

(3) *Symbols for units should be printed in lower case roman (upright) type. However, the symbol for a unit, derived from a proper name, should start with a capital roman letter, e.g., m (meter); A (ampere); Wb (weber); Hz (hertz).*

Prefixes—General Rules. (1) *The following prefixes should be used to indicate decimal fractions or multiples of a unit.*

$$\text{deci} \quad (=10^{-1}) \text{ d}$$

$$\text{centi} \quad (=10^{-2}) \text{ c}$$

$$\text{milli} \quad (=10^{-3}) \text{ m}$$

$$\text{micro} \quad (=10^{-6}) \mu$$

$$\text{nano} \quad (=10^{-9}) \text{ n}$$

$$\text{pico} \quad (=10^{-12}) \text{ p}$$

$$\text{femto} \quad (=10^{-15}) \text{ f}$$

$$\text{atto} \quad (=10^{-18}) \text{ a}$$

$$\text{kilo} \quad (=10^3) \text{ k}$$

$$\text{mega} \quad (=10^6) \text{ M}$$

$$\text{giga} \quad (=10^9) \text{ G}$$

$$\text{tera} \quad (=10^{12}) \text{ T}$$

(2) *The use of double prefixes should be avoided when single prefixes are available.*

$$\text{Not: } \text{m}\mu\text{s}, \quad \text{but: } \text{ns (nanosecond)}$$

$$\text{Not: } \text{kMW}, \quad \text{but: } \text{GW (gigawatt)}$$

$$\text{Not: } \mu\mu\text{F}, \quad \text{but: } \text{pF (picofarad)}$$

(3) *When a prefix is placed before the symbol of a unit, the combination of prefix and symbol should be considered as one new symbol, which can be squared or cubed without using brackets.*

EXAMPLES:

$$\text{cm}^2, \quad \text{mA}^2, \quad \text{s}^2$$

REMARK: cm^2 always means $(0.01 \text{ m})^2$ and never 0.01 m^2 .

Mathematical Operations. (1) *Multiplication of two units may be indicated in one of the following ways:*

$$\text{newton meter: } \text{Nm} \quad \text{N m} \quad \text{N} \cdot \text{m}$$

(2) *Division of one unit by another unit may be indicated in one of the following ways:*

$$\text{meter per second: } \frac{\text{m}}{\text{s}} \quad \text{m/s} \quad \text{m s}^{-1}$$

or by any other way of writing the product of m and s^{-1} . Not more than one solidus should be used.

EXAMPLES:

$$\text{Not: } \text{cm/s/s},$$

$$\text{but: } \text{cm/s}^2 = \text{cm s}^{-2}$$

$$\text{Not: } 1 \text{ poise} = 1 \text{ g/s/cm},$$

$$\text{but: } 1 \text{ poise} = 1 \text{ g/s} \cdot \text{cm} = 1 \text{ g s}^{-1} \text{ cm}^{-1}$$

$$\text{Not: } \text{J}^\circ\text{K/mol},$$

$$\text{but: } \text{J}^\circ\text{K mol} = \text{J}^\circ\text{K}^{-1} \text{ mol}^{-1}$$

Numbers and Figures. (1) *Numbers should be printed in upright type.*

(2) *Division of one number by another number may be indicated in the following ways:*

$$\frac{136}{273.15} \quad 136/273.15$$

or by writing it as the product of numerator and the inverse power of the denominator. In such cases, the number under the inverse power should always be placed between brackets.

REMARK: When the solidus is used and when there is any doubt where the numerator starts or the denominator ends, brackets should be used, as in the case of physical quantities.

(3) *To facilitate the reading of large numbers, the figures may be grouped in groups of three, but no comma should be used, since European convention uses the comma as a decimal point.*

EXAMPLE:

$$2\,573\,421\,736.01$$

Symbols for Chemical Elements, Nuclides, and Particles. (1) *Symbols for chemical elements* should be written in *roman* (upright) type. The symbol is not followed by a full stop.

EXAMPLES:

Ca C H He

(2) A nuclide is specified by the chemical symbol and the mass number, which should appear as a left superscript, e.g., ^{14}N . The atomic number may be shown too as a left subscript, e.g., ${}_7^{14}\text{N}$, if needed. The right superscript position may be used to indicate a state of ionization, e.g., Ca^{2+} , OH^- , or a state of excitation, e.g., $^{110}\text{Ag}^{\text{m}}$, $^4\text{He}^*$. The right subscript position is used to indicate the number of atoms of the specified nuclide or chemical element in a molecule, e.g., H_2SO_4 .

(3) Symbols for particles and quanta

neutron n

triton t

leptons e (electron), ν (neutrino), μ (muon)

mesons π (pion), K, η

baryons Λ , Σ , Ξ , Ω

proton p

α -particle α

deuteron d

photon γ

The charge of particles may be indicated by adding the superscript +, −, or 0.

EXAMPLES:

π^+ , π^- , η^0 , p^+ , p^- , e^+ , e^-

If in connection with the symbols p and e no charge is indicated, these symbols should refer to the positive proton and the negative electron respectively. The tilde above the symbol of a particle is often used to indicate the antiparticle of that particle (e.g., $\bar{\nu}$ for antineutrino).

Quantum States. (1) A symbol indicating the quantum state of a *system* such as an atom should be printed in capital roman (upright) type. The right subscript indicates the total angular momentum quantum number, and the left superscript indicates the multiplicity.

EXAMPLE:

$^2\text{P}_{3/2}$ (J = $\frac{3}{2}$, multiplicity: 2)

(2) A symbol indicating the quantum state of a single *particle* such as an electron should be printed in lower-case roman (upright) type. The right subscript may be used to indicate the total angular momentum quantum number of the particle in the case of j-j coupling.

EXAMPLE:

$p_{3/2}$ (electron state)

(3) The letter symbols corresponding to the angular momentum quantum number should be:

0 S, s 4 G, g 8 L, l
1 P, p 5 H, h 9 M, m

2 D, d 6 I, i 10 N, n

3 F, f 7 K, k 11 O, o

Nomenclature. (1) *Use of the word specific.* The word "specific" in English names for physical quantities should be restricted to the meaning "divided by mass."

EXAMPLES:

Specific volume volume/mass

Specific energy energy/mass

Specific heat capacity heat capacity/mass

(2) *Notation for covariant character of coupling:*

S Scalar coupling A Axial vector coupling

V Vector coupling P Pseudoscalar coupling

T Tensor coupling

(3) *Abbreviated notation for a nuclear reaction.* The meaning of the symbolic expression indicating a nuclear reaction should be the following:

$$\begin{array}{ccc} \text{Initial} & \left(\begin{array}{cc} \text{incoming} & \text{outgoing} \\ \text{particle} & \text{particle(s)} \\ \text{or} & \text{or} \\ \text{quantum} & \text{quanta} \end{array} \right) & \text{Final} \\ \text{nuclide} & & \text{nuclide} \end{array}$$

EXAMPLES:

$^{14}\text{N}(\alpha, p)^{17}\text{O}$ $^{59}\text{Co}(n, \gamma)^{60}\text{Co}$

$^{23}\text{Na}(\gamma, 3n)^{20}\text{Na}$ $^{31}\text{P}(\gamma, pn)^{29}\text{Si}$

(4) *Character of transitions.* Multipolarity of transition:

• Electric or magnetic	monopole	E0 or M0
" "	dipole	E1 or M1
" "	quadrupole	E2 or M2
" "	octupole	E3 or M3
" "	2 ⁿ pole	En or Mn

(5) *Nuclide.* A species of atoms, with specified atomic number and mass number should be indicated by the word *nuclide*, not by the word *isotope*.

Different nuclides having the same atomic number should be described as *isotopes*.

Different nuclides having the same mass number should be described as *isobars*.

Recommended Symbols for Physical Quantities.

REMARK: (1) Where several symbols are given for one quantity, and no special indication is made, they have equal weight.

(2) Only symbols are given for units in the tables below. All symbols are given in the mksA system of units. Units having special names in this system will be found listed separately on p. 702, together with their international symbols.

(3) Decimal multiples and submultiples of units are not explicitly mentioned in the tables; this does not mean, however, that they are not recommended.

Table of Symbols for Physical Quantities

Quantity	Symbol	International Symbol for Unit
SPACE AND TIME		
length	l	m
breadth	b	m
height	h	m
radius	r	m
diameter: $d = 2r$	d	m
path: $L = \int ds$	L, s	m
area	A, S	m ²
volume	V, v	m ³
plane angle	$\alpha, \beta, \gamma, \theta, \vartheta, \phi$	rad
solid angle	ω, Ω	sr
wave length	λ	m
wave number: $\sigma = 1/\lambda$	σ	m ⁻¹
circular wave number: $k = 2\pi/\lambda$	k	m ⁻¹
time	t	s
period	T	s
frequency: $\nu = 1/T$	ν, f	Hz
angular frequency: $\omega = 2\pi\nu$	ω	rad/s*
velocity: $v = ds/dt$	c, u, v	m/s
angular velocity: $\omega = d\phi/dt$	ω	rad/s
acceleration: $a = dv/dt$	a	m/s ²
angular acceleration: $\alpha = d\omega/dt$	α	rad/s ²
gravitational acceleration	g	m/s ²
standard gravitational acceleration	g_n	m/s ²
relative velocity: v/c	β	...
MECHANICS		
mass	m	kg
density: $\rho = m/V$	ρ	kg/m ³
reduced mass	μ, μ_r	kg
momentum: $p = mv$	p, p	kg·m/s
moment of inertia: $I = \int r^2 dm$	I, J	kg·m ²
force	F, f	$N \left(= \frac{\text{kg} \cdot \text{m}}{\text{s}^2} \right)$
weight	G, W	N
moment of force	M, M	N·m
pressure	p	N/m ²
normal stress	σ	N/m ²
shear stress	τ	N/m ²
gravitational constant: $F(r) = Gm_1m_2/r^2$	G	N·m ² /kg ²
modulus of elasticity, Young's modulus: $\sigma = E\Delta l/l$	E	N/m ²
shear modulus: $\tau = Gtg\gamma$	G	N/m ²
compressibility: $\kappa = -(1/V)dV/dp$	κ	m ² /N
bulk modulus: $K = 1/\kappa$	K	N/m ²
viscosity	η	N·s/m ²
kinematic viscosity: $\nu = \eta/\rho$	ν	m ² /s ²
friction coefficient	f	...
surface tension	γ, σ	N/m
energy	E, U	$J \left(= \frac{\text{kg} \cdot \text{m}^2}{\text{s}^2} \right)$
potential energy	V, E_p	J
kinetic energy	T, E_k	J
work	W, A	J
power	P	W (= J/s)
efficiency	η	...
Hamiltonian function	H	J
Lagrangian function	L	J
relative density	d	...

* The international bodies give s⁻¹ as the unit for angular frequency.

Quantity	Symbol	International Symbol for Unit
MOLECULAR PHYSICS		
number of molecules	N	
number density of molecules: $n = N/V$	n	m^{-3}
Avogadro's constant	$N_0, (L)$	mol^{-1}
molecular mass	m	kg
molecular velocity vector with components	$\mathbf{c}, (c_x, c_y, c_z)$	m/s
	$\mathbf{u}, (u_x, u_y, u_z)$	m/s
molecular position vector with components	$\mathbf{r}, (x, y, z)$	m
molecular momentum vector with components	$\mathbf{P}, (p_x, p_y, p_z)$	$\text{kg} \cdot \text{m/s}$
average velocity	$\mathbf{c}_0, \mathbf{u}_0, \bar{c}, \bar{u}$	m/s
most probable speed	\bar{c}, \bar{u}	m/s
mean free path	l	m
molecular attraction energy	ε	J
interaction energy between molecules i and j	ϕ_{ij}, V_{ij}	J
velocity distribution function: $n = \int f \, dc_x dc_y dc_z$	$f(c)$...
generalized coordinate	q	...
generalized momentum	p	...
volume in γ phase space	Ω	...
Boltzmann's constant	k	J/K
$1/kT$ in exponential functions	β	J^{-1}
gas constant per mole	R	$\text{J/mol} \cdot \text{K}$
partition function	Q, Z	...
diffusion coefficient	D	m^2/s
thermal diffusion coefficient	D_T	m^2/s
thermal diffusion ratio	k_T	...
thermal diffusion factor	α_T	...
characteristic temperature	Θ	K
Debye temperature: $\Theta_D = h\nu_D/k$	Θ_D	K
Einstein temperature: $\Theta_E = h\nu_E/k$	Θ_E	K
rotational temperature: $\Theta_r = h^2/8\pi^2Ik$	Θ_r	K
vibrational temperature: $\Theta_v = h\nu/k$	Θ_v	K
THERMODYNAMICS		
quantity of heat	Q	J
work	W, A	J
temperature	$t, (\theta)$	$^\circ\text{K}$
thermodynamic temperature	$T, (t)$	$^\circ\text{K}$
entropy	S	J/K
internal energy	U	J
Helmholtz function, free energy: $F = U - TS$	F	J
enthalpy: $H = U + pV$	H	J
Gibbs function: $G = U + pV - TS$	G	J
linear expansion coefficient	α	$^\circ\text{K}^{-1}$
cubic expansion coefficient	γ	$^\circ\text{K}^{-1}$
thermal conductivity	λ	$\text{W/m} \cdot \text{K}$
specific heat capacity	c_p, c_v	$\text{J/kg} \cdot \text{K}$
molar heat capacity	C_p, C_v	$\text{J/mol} \cdot \text{K}$
Joule-Thomson coefficient	μ	$^\circ\text{K} \cdot \text{m}^2/\text{N}$
ratio of specific heats	k, γ	...
ELECTRICITY AND MAGNETISM		
quantity of electricity	Q	$\text{C} (= \text{A} \cdot \text{s})$
charge density	ρ	C/m^3
surface charge density	σ	C/m^2
electric potential	V, Φ	$\text{V} (= \text{W/A})$
electric field	\mathbf{E}, E	$\text{N/C}, \text{V/m}$
electric displacement	\mathbf{D}, D	C/m^2
capacitance	C	$\text{F} (= \text{C/V})$
permittivity: $\varepsilon = D/E$	ε	F/m
permittivity of vacuum	ε_0	F/m
relative permittivity: $\varepsilon_r = \varepsilon/\varepsilon_0$	ε_r	...

Quantity	Symbol	International Symbol for Unit
dielectric polarization: $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$	\mathbf{P}, P	C/m ²
electric susceptibility	χ_e	...
polarizability	α, γ	C · m ² /V
electric dipole moment	\mathbf{p}, p	C · m
electric current	I	A
electric current density	\mathbf{J}, J	A/m ²
magnetic field	\mathbf{H}, H	A/m
magnetic induction	\mathbf{B}, B	T (= Wb/m ²)
magnetic flux	Φ	Wb (= V · sec)
permeability: $\mu = B/H$	μ	H/m
permeability of vacuum	μ_0	H/m
relative permeability: $\mu_r = \mu/\mu_0$	μ_r	...
magnetization: $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$	\mathbf{M}, M	A/m
magnetic susceptibility	χ_m	...
electromagnetic moment	μ, μ_e, m, m	A · m ²
magnetic polarization: $\mathbf{B} = \mu_0 \mathbf{H} + \mathbf{J}$	\mathbf{J}	Wb/m ²
magnetic dipole moment	\mathbf{j}, J	Wb · m
resistance	R	Ω (= V/A)
reactance	X	Ω
impedance: $Z = R + iX$	Z	Ω
admittance: $Y = 1/Z = G + iB$	Y	Ω^{-1}
conductance	G	Ω^{-1}
susceptance	B	Ω^{-1}
resistivity	ρ	$\Omega \cdot m$
conductivity: $1/\rho$	γ, σ	($\Omega \cdot m$) ⁻¹
self-inductance	L	H (= V · s/A)
mutual inductance	M, L_{12}	H
phase number	m	...
loss angle	δ	rad
number of turns	N	...
power	P	W
Poynting vector	\mathbf{S}, S	W/m ²
vector potential	\mathbf{A}	Wb/m

LIGHT, RADIATION

quantity of light	$Q(Q_v)$	lm · s
luminous flux	$\Phi(\Phi_v)$	lm
luminous intensity	$I(I_v)$	cd
illuminance	$E(E_v)$	lx (= lm/m ²)
luminance	$L, (L_v)$	cd/m ²
luminous emittance	M	lm · m ²
radiant energy	$Q, W, (Q_e)$	J
radiant flux	$\Phi, (P, \Phi_e)$	W
radiant intensity	$I, (I_e)$	W/sr
irradiance	$E, (E_e)$	W/m ²
radiance	$L, (L_e)$	W/sr · m ²
radiant emittance	$M, (M_e)$	W/m ²
absorptance	$\alpha(\lambda)$...
reflectance	$\rho(\lambda)$...
transmittance	$\tau(\lambda)$...
linear attenuation coefficient	$\mu(\lambda)$	m ⁻¹
linear absorption coefficient	$a(\lambda)$	m ⁻¹
refractive index	n	...
Stefan-Boltzmann constant: $\sigma = T^4/M$	σ	W/m ² · °K ⁴
luminous efficacy: Φ_v/Φ_e	K	lm/W
luminous efficiency: K/K_{\max}	V	...

ACOUSTICS

velocity of sound	c	m/sec
sound energy flux	P	W
reflection factor	ρ	...
acoustic absorption factor: $1 - \rho$	$\alpha, (\alpha_a)$...

Quantity	Symbol	International Symbol for Unit
transmission factor	τ	...
dissipation factor: $\alpha = \tau$	δ	...
loudness level	$L_N, (\Delta)$	phon
reverberation time	T	s
specific acoustic impedance	$Z_n, (W)$	$N \cdot s/m^3$
acoustic impedance	$Z_a, (Z)$	$N \cdot s/m^5$
mechanical impedance	$Z_m, (\omega)$	$N \cdot s/m$

ATOMIC AND NUCLEAR PHYSICS

atomic number, proton number	Z	...
mass number	A	...
neutron number: $N = A - Z$	N	...
elementary charge	e	C
electron mass	m, m_e	kg
proton mass	m_p	kg
neutron mass	m_n	kg
meson mass	m_π, m_μ	kg
atomic mass	$m_a, m(X)$	kg
(unified) atomic mass constant: $m_u = m_a(^{12}\text{C})/12$	m_u	u
magnetic moment of atom or nucleus	μ	$A \cdot m^2$
magnetic moment of proton	μ_p	$A \cdot m^2$
magnetic moment of neutron	μ_n	$A \cdot m^2$
magnetic moment of electron	μ_e	$A \cdot m^2$
Bohr magneton	μ_B	$A \cdot m^2$
Planck constant $\left(\frac{h}{2\pi} = \hbar\right)$	h	J · s
principal quantum number	n, n_1	...
orbital angular momentum quantum number	L, l_1	...
spin quantum number	S, s_1	...
total angular momentum quantum number	J, j_1	...
magnetic quantum number	M, m_1	...
nuclear spin quantum number	I	...
hyperfine quantum number	F	...
rotational quantum number	J, K	...
vibrational quantum number	v	...
quadrupole moment	Q	m^2
Rydberg constant	R_∞	m^{-1}
Bohr radius: $a_0 = 4\pi\epsilon_0\hbar^2/m_e e^2$	a_0	m
fine structure constant: $\alpha = e^2/4\pi\epsilon_0\hbar c$	α	...
mass excess: $m_a - Am_u$	Δ	kg
mass defect	B	kg
packing fraction: Δ/Am_u	f	...
nuclear radius: $R = R_0 A^{1/3}$	R	m
nuclear magneton	μ_N	$A \cdot m^2$
g -factor of nucleus: $\mu = gI\mu_N$	g	...
gyromagnetic ratio: $\gamma = \mu/I\hbar$	γ	$A \cdot m^2/J \cdot s$
Larmor (angular) frequency	ω_L	rad/s
cyclotron (angular) frequency	ω_c	rad/s
level width	Γ	J
mean life	τ	s
reaction energy	Q	J
cross section	σ	m^2
macroscopic cross section	Σ	m^2
impact parameter	b	m
scattering angle	$\vartheta, \theta, \varphi$	rad
particle flux density	Φ	$s^{-1} \cdot m^{-2}$
particle fluence	Φ	m^{-2}
energy flux density	ψ	W/m^2
energy fluence	F	J/m^2
internal conversion coefficient	α	...
reaction energy	Q	J
half-life	$T_{1/2}$	s

Quantity	Symbol	International Symbol for Unit
decay constant, disintegration constant	λ	s^{-1}
activity	A	s^{-1}
Compton wavelength: $\lambda_c = h/mc$	λ_c	m
electron radius: $r_e = e^2/mc^2$	r_e	m
linear attenuation coefficient	μ, μ_1	m^{-1}
atomic attenuation coefficient	μ_a	m^2
mass attenuation coefficient	μ_m	m^2/kg
linear stopping power	S, S_1	J/m
atomic stopping power	S_a	$J \cdot m^2$
linear range	R, R_1	m
recombination coefficient	α	m^3/s
linear ionization by a particle	N_{i1}	m^{-1}
total ionization by a particle	N	...
number of neutrons per fission	ν	...
number of produced neutrons per absorption	η	...
absorbed dose	D	J/kg (rad)
linear energy transfer	L	J/m

CHEMICAL PHYSICS

amount of substance	n, ν	mol
molar mass of substance B	M_B	kg/mol
molar concentration of substance B	c_B	mol/m ³
mole fraction of substance B	x_B	...
molar internal energy	$U_m, (E_m)$	J/mol
mass fraction of substance B	w_B	...
molality of solution	m	mol/kg
chemical potential	μ	J/mol
activity of component B	λ_B	...
relative activity	a_B	...
activity coefficient (molar fraction, molality, molarity scales)	f_B, γ_B, γ_B	...
osmotic pressure	Π	N/m ²
osmotic coefficient	g, ϕ	...
stoichiometric number of molecule B	ν_B	...
affinity	A	J/mol
equilibrium constant	K_p	(N/m ²)
charge number of ion	z	...
Faraday constant	F	C/mol
ionic strength	I	mol/kg
fugacity of component B	f_B	N/m ²
electrolytic conductivity	σ, κ, γ	$\Omega^{-1} \cdot m^{-1}$
transport number	t	...
degree of dissociation	α	...

Recommended Mathematical Symbols:

GENERAL SYMBOLS

equal to	$=$
not equal to	\neq, \neq
identically equal to	\equiv
corresponds to	\sim, \sim
approximately equal to	\approx
proportional to	\propto, \propto
approaches	\rightarrow, \rightarrow
larger than	$>, >$
smaller than	$<, <$
much larger than	\gg
much smaller than	\ll
larger than or equal to	\geq, \geq, \geq
smaller than or equal to	\leq, \leq, \leq
plus	$+$
minus	$-$

Quantity	Symbol	International Symbol for Unit
plus or minus	\pm	
a raised to the power n	a^n	
magnitude of a	$ a $	
square root of a	$\sqrt{a}, \sqrt[3]{a}, a^{1/2}$	
mean value of a	$\bar{a}, \langle a \rangle, \langle a \rangle_{av}$	
factorial p	$p!$	
binomial coefficient: $n!/p!(n-p)!$	$\binom{n}{p}$	

LETTER SYMBOLS AND LETTER EXPRESSIONS FOR MATHEMATICAL OPERATIONS

These should be written in roman (or upright) type	
exponential of x	$\exp x, e^x$
base of natural logarithms	e
logarithm to the base a of x	$\log_a x$
natural logarithm of x	$\ln x$
common logarithm of x	$\lg x, \log x$
summation	Σ
product	Π
finite increase of x	Δx
variation of x	δx
total differential of x	dx
function of x	$f(x), f(x)$
limit of $f(x)$	$\lim f(x)$

TRIGONOMETRIC FUNCTIONS

sine of x	$\sin x$
cosine of x	$\cos x$
tangent of x	$\tan x, \operatorname{tg} x$
cotangent of x	$\cot x, \operatorname{ctg} x$
secant of x	$\sec x$
cosecant of x	$\operatorname{cosec} x$

REMARKS: (a) It is recommended to use for the *inverse circular functions* the symbolic expressions for the corresponding circular function preceded by the letters: arc.

EXAMPLES:

$\arcsin x, \arccos x, \arctan x$ or $\operatorname{arctg} x$, etc.

Sometimes the notation $\sin^{-1} x, \tan^{-1} x$, etc., is used.

(b) It is recommended to use for the *hyperbolic functions* the symbolic expressions for the corresponding hyperbolic function, followed by the letter: h.

EXAMPLES:

$\sinh x, \cosh x, \tanh x$ or $\operatorname{tgh} x$, etc.

(c) It is recommended to use for the *inverse hyperbolic functions* the symbolic expression for the corresponding hyperbolic function preceded by the letters: ar.

$\operatorname{arsinh} x, \operatorname{arcosh} x$, etc.

COMPLEX QUANTITIES

imaginary unit ($i^2 = -1$)	i, j
real part of z	$\operatorname{Re} z, z'$
imaginary part of z	$\operatorname{Im} z, z''$
modulus of z	$ z $
argument of z : $z = z \exp i\phi$	$\arg z, \varphi$
complex conjugate of z , conjugate of z	z^*

REMARK: Sometimes the notation \bar{z} is used for the complex conjugate of z .

VECTOR CALCULUS (SEE ALSO P. 695)

absolute value	$ A , A$
differential vector operator	$\partial/\partial r, \nabla$
gradient	$\operatorname{grad} \varphi, \nabla \varphi$
divergence	$\operatorname{div} A, \nabla \cdot A$
curl	$\operatorname{curl} A, \operatorname{rot} A, \nabla \times A$
Laplacian	$\Delta \varphi, \nabla^2 \varphi$
d'Alembertian	$\square \varphi$

MATRIX CALCULUS

transpose of matrix A	$\tilde{A}_{ij} = A_{ji}, \tilde{A}$
complex conjugate of A	$(A^*)_{ij} = (A_{ij})^*, A^*$
Hermitian conjugate of A	$(A^\dagger)_{ij} = A_{ji}^*, A^\dagger$

International Symbols for Units. Unit systems.

(1) A *coherent system of units* is a system based on a certain set of "basic units" from which all "derived units" are obtained by multiplication or division without introducing numerical factors. In addition there are "dimensionless units," in particular the radian, symbol: rad, for plane angle and the steradian, symbol: sr, for solid angle.

(2) The *cgs system* or *cm-g-s system* is a coherent system of units based on *three basic units* for the three basic quantities length, mass and time:

centimeter	cm
gram	g
second	s

In the field of *mechanics* the following units of this system have special names and symbols, which have been approved by the General Conference on Weights and Measures:

l, b, h	centimeter	cm
t	second	s
m	gram	g
f, ν	hertz ($=s^{-1}$)	Hz
F	dyne ($=g \cdot cm/s^2$)	dyn
E, U, W, A	erg ($=g \cdot cm^2/s^2$)	erg
p	microbar ($=dyn/cm^2$)	μ bar
η	poise ($=dyn \cdot s/cm^2$)	P

In the field of *electricity and magnetism* several variants of the cgs unit system have been developed, in particular the *electrostatic cgs system* and *electromagnetic cgs system*. Special names and symbols for some of the units of the second system are:

H	oersted ($=cm^1 \cdot g^1 \cdot s^{-1}$)	Oe
B	gauss ($=cm^1 \cdot g^1 \cdot s^{-1}$)	G
Φ	maxwell ($=cm \cdot g^1 \cdot s^{-1}$)	Mx

(3) The *mksA system* or m-kg-s-A system is a coherent system of units for mechanics, electricity and magnetism, based on *four basic units* for the four basic quantities length, mass, time, and electric current intensity:

meter	m
kilogram	kg
second	s
ampere	A

REMARK: The system based on these four units was given the name *Giorgi system* by the International Electrotechnical Committee in 1958. The mechanical system, which is based on the first three units only, has the name *mks system*.

The following units of the mksA system have special names and symbols, which have been approved by the General Conference on Weights and Measures:

l, b, h	meter	m
t	second	s
m	kilogram	kg
ν, f	hertz ($=s^{-1}$)	Hz
F	newton ($=kg \cdot m/s^2$)	N
E	joule ($=kg \cdot m^2/s^2$)	J
P	watt ($=J/s$)	W
I	ampere	A
Q	coulomb ($=A \cdot s$)	C
V	volt ($=W/A$)	V
C	farad ($=C/V$)	F
R	ohm ($=V/A$)	Ω
L	henry ($=V \cdot s/A$)	H
Φ	weber ($=V \cdot s$)	Wb
B	tesla ($=Wb/m^2$)	T

(4) In the field of *thermodynamics* one introduces an additional basic unit, corresponding to the basic quantity: *thermodynamic temperature*, the unit being the degree Kelvin, symbol: $^{\circ}K$.

When the *customary temperature* is used, defined by $t = T - T_0$, where $T_0 = 273.15^{\circ}K$, this is usually expressed in degrees Celsius, symbol: $^{\circ}C$. For *temperature interval* the name degree, symbol: deg, is often used, the indications "Kelvin" or "Celsius," indicating the zero-point of the temperature scale used, being irrelevant in this case.

(5) In the field of *photometry* one introduces an additional basic unit, corresponding to the basic quantity: *luminous intensity*, this unit being the candela, symbol: cd. Special names for units in this field are:

I	candela	cd
Φ	lumen	lm
E	lux ($=lm/m^2$)	lx

(6) *The International System of Units*. For the system based on the six basic units:

meter	m	ampere	A
kilogram	kg	degree Kelvin	$^{\circ}K$
second	s	candela	cd

The name *International System of Units* has been recommended by the Conférence Générale des Poids et Mesures in 1960:

(7) In the field of *chemical and molecular physics*, in addition to the basic quantities defined above, *amount of substance* is also treated as a basic quantity. The basic recommended basic unit is the mole, symbol: mol. The mole is defined as the amount of substance, which contains the same number of elementary units (e.g. atoms, molecules, ions, electrons, or groups of such entities corresponding to a stated formula), as there are atoms in exactly 12 grams of the pure carbon nuclide ^{12}C .

INCOHERENT UNITS

l	ångström	Å
σ	barn ($=10^{-28} cm^2$)	b
V	liter	l
t, τ, T_1	minute	min
t, τ, T_1	hour	h
t, τ, T_1	day	d
t, τ, T_1	year	a
p	atmosphere	atm
P	kilowatt-hour	kWh
Q	calorie	cal
Q	kilocalorie	kcal
E, Q	electronvolt	eV
m	ton ($=1000 kg$)	t
m_A, m	(unified) atomic mass unit	u
p	bar ($=10^6 dyn/cm^2$)	bar

REMARK: The (unified) atomic mass unit is defined as $\frac{1}{12}$ th of the mass of an atom of the ^{12}C nuclide.

HUGH C. WOLFE
PAUL J. KLIAUGA

Cross-references: CONSTANTS, FUNDAMENTAL.

SYMMETRY. See CONSERVATION LAWS AND SYMMETRY.

SYNCHROTRONS

The term synchrotron has come to mean "a ring shaped device for accelerating charged particles, e.g., electrons, protons, deuterons, to relativistic energies by the repeated passage of the particles, at essentially constant radius, through a time-varying electric field which alternates in direction at a fixed or variable frequency." Special examples are: (1) weak focusing synchrotrons; (2) alternating-gradient synchrotrons (AGS); (3) fixed-field alternating-gradient synchrotrons (FFAG), or azimuthally varying field synchrotrons (AVF). All these machines have evolved from the early accelerators invented by E. O. Lawrence (cyclotron—1930), D. W. Kerst (betatron—1940), the synchronous acceleration principle discovered by E. M. McMillan and V. Veksler (1945), and the alternating-gradient strong focusing principle of Christofilos (1950), and Livingston, Courant and Snyder (1952).

There are basically six components or functions which a synchrotron must provide: (1) a source of particles, e.g., electrons or protons with energy and direction suitable for injection; (2) a vacuum chamber; (3) a magnet to bend particles in a circle; (4) focusing of particles; (5) acceleration of particles; (6) ejection of particles.

Ion Source. Electrons or ions to be accelerated are generally produced in an external device by a hot cathode or a gaseous discharge and then, after being electromagnetically focused into a narrow pencil, they are accelerated up to several tens of MeV before injection into the synchrotron magnet gap. The optimum injector energy depends upon a number of factors, e.g., minimum magnet field

strength at which the magnetic field shape will produce satisfactory focusing, space charge limitations, design of the radio-frequency acceleration system, and "inflection" problems associated with bending the particles into orbit. Generally speaking, a fairly high injector energy is preferred since most of the above problems are eased as the injector energy increases. Very close tolerances must be placed on the energy and angular spread of the injected beam; otherwise the particles will strike the walls of the vacuum chamber after only a few turns. Injector accelerators generally in use are: (1) Cockroft Walton (500 kV); (2) Van de Graaff (3 to 5 MeV); (3) linear accelerators (15 to 50 MeV).

Vacuum Chamber. Since particles are scattered by collision with air molecules, it is necessary to provide a good vacuum over the entire region traversed by the particles while they are undergoing acceleration. The most critical period is when the particles are at low velocity, for then the Rutherford nuclear scattering and small angle atomic scattering are most probable. A vacuum of 10^{-6} mm of Hg is generally quite sufficient to reduce gas scattering losses to negligible proportions. In order that the vacuum chamber walls should not interfere with the magnetic field, or its space and time derivatives, it is customary to employ either insulating walls, e.g., ceramic, epoxy fiber glass, or laminated metal structures made vacuum tight by an outer vacuum envelope of rubber or epoxy fiber glass.

Bending. A particle of rest mass m_0 , charge $Z \cdot e$, and kinetic energy T , moving in a direction perpendicular to a magnetic field of B gauss, will move in a circle of radius R given by:

$$B \cdot R = \frac{1}{9.15Z} (T^2 + 2TE_0)^{1/2} \quad \begin{matrix} \text{(kilogauss,} \\ \text{feet, MeV)} \end{matrix}$$

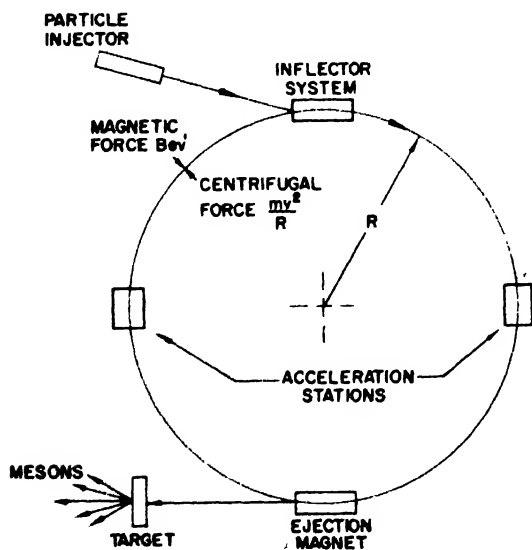


FIG. 1. Synchrotron.

where $E_0 = m_0c^2$, the rest energy of the particle. From this equation it can be seen that as the particle energy is increased from a very low initial energy to the final energy of several thousand MeV, the product $B \cdot R$ must increase manifold. If B is held constant in time during the acceleration cycle and if the space average of B is also relatively constant, then R must increase with energy thus leading to the need for a very large radial aperture over which the field must be maintained. This is the case with cyclotrons. The radial aperture can be drastically reduced in either of two ways: (1) time modulate (e.g., pulse) the magnetic field so that it increases from a few tens or hundreds of gauss at ion injection time up to 10 to 20 kilogauss at which field the maximum particle energy is reached; (2) shape the magnetic field such that it increases sharply with increasing radius. The latter approach, while seemingly obvious, actually requires a very sophisticated angular and radial variation of the field in order to retain focusing in the direction parallel to the magnetic field. (It has been done by L. H. Thomas, 1938; Ohkawa, 1955; Kolomenskij, Petukhov,

and Rabinovitch, 1955; Symon, 1955.) Accelerators based on this time-independent but spatially varying field are generally referred to as FFAG or Spiral Sector Ring Accelerators. Their major utility appears to be in the field of high ion currents, which they can achieve by virtue of their non-pulsed character, but they appear to be practicable up to only a few thousand MeV of energy since they still require considerable radial aperture. By far the most widely exploited type of synchrotron is that first mentioned, i.e., the time-varying magnetic field, which holds the particles at an essentially constant radius while gaining energy. This technique, which leads to a simpler magnet design and a minimum of weight, will be described in detail in the remainder of this article.

Focusing. A magnetic field which is perfectly uniform over the entire circular path of the particle and over the radial aperture leads to stable motion only for radial displacements of the particle. Particles displaced parallel to B (i.e., out of the orbit plane) are not refocused back on to the original circle and are quickly lost to the chamber walls. This can be readily understood by noting that motion parallel to B generates no $eV \times B$ force, and therefore such motion persists until the particle strikes the chamber walls. However, by deliberately introducing curved lines of force, it is possible to produce $eV \times B$ forces whose net effect results in restoring forces for all displacements from the ideal circular orbit. Subsequent to the pioneer cyclotron and

betatron work, in which a simple *decrease* of magnetic field with *increasing* radius leads to a weak-focusing action, there has evolved a wide variety of field shapes which have much stronger focusing action. All these strong-focusing fields are characterized by rapid spatial changes in the field strength, reversal of sign of field gradients and even of the field itself. The most widely employed alternating gradient scheme is one in which a magnet sector that focuses *radially*, but defocuses *axially*, is followed by a sector which has the reverse property, i.e., *defocuses* radially and focuses axially. The net effect is strong focusing in *both* directions. There is a close analogy between alternating-gradient synchrotron focusing and an alternating series of convergent and divergent optical lenses whose net action is focusing. As a result of this arrangement of magnetic lenses, it is possible to use larger field gradients than in the weak-focusing case and thereby to achieve very strong focusing action. For example, an alternating-gradient 35 000-MeV accelerator, 560 feet in diameter, requires a vacuum chamber with a cross section of only 2.7×6 inches, whereas a weak-focusing, constant-gradient synchrotron would require at least ten times more aperture for equivalent performance.

Acceleration. The increase in particle energy actually takes place when the circling charged particles pass through electric fields spaced around the circumference of the magnet. These fields must be time varying since a complete, repetitive traversal of a static field leads to zero energy

TABLE 1. SOME TYPICAL SYNCHROTRONS

	Princeton-Pennsylvania Accelerator, Princeton, N.J.	Brookhaven National Laboratory, Upton, L. I.	Joint Institute for Nuclear Research, Dubna, USSR	CEA Electron Accelerator, Cambridge, Mass.
1. Focusing scheme	Weak	Strong	Weak	Strong
2. Particle	Protons	Protons	Protons	Electrons
3. Maximum energy, BeV	3.0	33.0	10	6
4. Pulse rate, per min	1140	30	5	3600
5. Particles per pulse	$7 \cdot 10^{10}$	$5 \cdot 10^{11}$	$5 \cdot 10^{10}$	$4 \cdot 10^{10}$
6. Particles per second	$1.4 \cdot 10^{12}$	$2.5 \cdot 10^{11}$	$4 \cdot 10^9$	$2.4 \cdot 10^{12}$
7. Orbit radius, ft	40	421.40	91.6	86.5
8. Magnet weight, ton	400	4400	36,460	298
9. Magnet power, kW	1200 Average	$30 \cdot 10^3$ (peak)	$140 \cdot 10^3$ (peak)	1100 (average)
10. Vacuum chamber, in.	$7 \cdot 2.5$	$6 \cdot 2.5$	$59 \cdot 14$	$5.2 \cdot 1.5$
11. Acceleration system	4 drift tubes 4 ferrite cavities	12 double ferrite cavities	2 drift tubes	16 double cavities
12. Acceleration frequency, Mc/sec	2.5 - 30.0	1.4 - 4.5	0.182 - 1.45	475.8
13. rf cycles per revolution	8	12	1	360
14. Energy gain per turn, keV	60 (max)	100	2.5	4.5 MeV radiation loss per turn
15. rf power, peak kW	250	240	?	300
16. rf power, ave kW	125	120	?	80
17. Injector, MeV	Van de Graaff	Linac	Linac	Linac
	3.0	50	9	28
18. Magnetic field (injection), gauss	275	121	150	25.4
19. Magnetic field (max), kilogauss	14	13	13	7.6

gain. The usual arrangement is one in which the acceleration electrodes are excited by a sinusoidal voltage whose frequency is a harmonic (one to several hundred) of the particle rotation frequency and whose amplitude is such that particles can gain sufficient energy to match the rising magnetic field. Because of the principle of synchronous phase stability, discovered by McMillan and Veksler, particles with a wide spread in phase angle relative to the radio frequency are captured in stable "buckets" in phase space and are accelerated at just the correct rate, on the average, to stay in the middle of the vacuum chamber. Of course, the frequency of the accelerating field must be steadily and precisely increased as the particle rotation frequency increases. In some weak-focusing synchrotrons, this is one of the most difficult engineering tasks.

Ejection. When the particles have been accelerated to full energy, they are caused either to run into an internal target or they are ejected by one of several schemes which usually involve exciting strong radial betatron oscillations with the last oscillation carrying the particles into an ejection magnet which deflects the beam clear of the synchrotron magnet structure.

In Table I are listed the important parameters of several typical existing synchrotrons. The

principles discussed above for the alternating-gradient, strong-focusing accelerator (AGS) are believed to be capable of extension to almost any energy, limited only by financial considerations. There are now serious proposals before the Government agencies for synchrotrons with energy in the 200 to 1000 BeV range. The latter would cost around 700 million dollars and would be 4452 meters in diameter.

MILTON G. WHITE

References

- Livingood, J. J., "Cyclic Particle Accelerators," Princeton, N.J., D. Van Nostrand Co., 1961.
- Livingston, M. S., and Blewett, J. P., "Particle Accelerators," New York, McGraw-Hill Book Co., 1962.
- Wilson, R. R., and Littauer, R., "Accelerators, Machines of Nuclear Physics," Science Study Series, Garden City, N.Y., Doubleday Anchor, 1960.
- Green, G. K., and Courant, E. D., in F. Flugge, Ed., "Handbuch der Physik," Vol. 44, Berlin, Springer Verlag, pp. 218-340, 1959.

Cross-references: ACCELERATORS, LINEAR; ACCELERATORS, PARTICLE; BETATRON; CYCLOTRON; IONIZATION; ACCELERATOR, VAN DE GRAAFF.

T

TELEGRAPHY

Modern telegraphy has evolved into a science of such electronic sophistication that it bears practically no resemblance to its technological genesis. New developments arise so rapidly that today's theory quite literally becomes tomorrow's practice. Also, the scope of telegraphy has been broadened to include many new record services, such as data, facsimile, telemetry, etc.

Basically, most written telegraphed intelligence begins with a teleprinter, a typewriter-like piece of equipment, which performs the functions of coding, sending, receiving and decoding telegraph signals using only make-and-break transmission. Synchronism is achieved by the "start-stop" method by which a "start" pulse releases corresponding parts of the printing and transmitting units simultaneously. These parts operate in unison until the "stop" pulse arrests their motion and holds them until the succeeding start pulse arrives. The whole operation takes only 154 msec, which is approximately 390 characters or 65 words per minute. Faster speeds up to 100 words per minute are used where it is required.

A network of 15 high-speed switching centers, located throughout the country, cuts the manual handling of telegrams to one simple typing operation at the point of origin. The originating operators easily route outgoing telegrams to other switching centers by transmitting two "call letters" at the beginning of messages. An "electrical brain" in the connecting switching center is activated by these call letters and selects the proper circuit, automatically speeding the telegram to its destination.

Incoming telegrams, addressed to a point within the switching area, are received on a telegraph device known as a printer-perforator, which simultaneously types the telegram on the top of the tape while perforating code combinations in the tape. As each telegram arrives, a switching clerk presses a button marked with the name of its destination, thus connecting the message with a wire and automatically sending the telegram to its destination.

Western Union uses carrier systems, which multiply by many times the capacity of its wire and radio beam circuits, to make available the channels required to interconnect its country-wide network of switching centers. The fundamental principle of the carrier system is that

alternating or pulsating currents of a number of different frequencies are passed over a circuit, each frequency serving as a transmission channel. Individual transmitting devices control the application of the current at each frequency, and at the receiving end, the complex current composed of all the various frequencies is passed through electrical filters. Each filter segregates a particular frequency, which in turn is transmitted to appropriate receiving devices. In this way, a large number of individual channels may be obtained from one electrical circuit.

Since 1946 Western Union has been transmitting over a radio beam system linking New York and Philadelphia. The beam system was extended to Washington, D.C., and Pittsburgh two years later, and extensions to other cities followed. As originally designed, this network can transmit 2000 telegrams simultaneously in both directions.

The new 7500-mile transcontinental microwave system is now providing greatly enlarged capacity accommodating all known forms of communications at extremely high speeds and in large volume. It will handle sophisticated computer "talk," high-speed data and facsimile, and voice—as well as conventional telegraph. Engineered to allow for 100 per cent expansion, the network will be capable of beaming 12,000 simultaneous, two-way messages.

Towers in the radio beam system are spaced at intervals of 30 to 50 miles in "line of sight," as microwaves travel in a straight line. The towers are necessary to redirect the beam at these distances to compensate for the curvature of the earth. Super high-frequency radio waves are transmitted by directional beams from tower to tower where unattended stations automatically amplify and instantaneously relay the signals.

Atmospheric static is not a problem in the microwave region, and radio beam transmission is not bothered by electrical disturbances. Concentration of the available power into a sharply defined directional beam makes it possible to utilize the same radio-frequency channel over and over again.

The carrier system used on the radio beam furnishes 600 independent voice circuits and occupies the frequency range from 60 kc to 2.54 Mc. Each voice band provides a working transmission path extending from 200 to 3600 cycles, sufficient to accommodate 20 narrow-band

(150-cycle spacing) telegraph carrier channels or ten 180-baud data channels.

By supplementary equipment arrangement, the system may be modified to provide frequency bands roughly equivalent to 2, 4, 8 or 12 voice bands for use in broadband switching applications.

Telegraph facsimile has been so simplified in design and operation that today it is growing in importance as a means of fast, efficient communication. At the sending point, matter to be transmitted is placed upon the revolving drum of a facsimile machine and scanned by a finely focused beam of light. The minute gradations of current intensity thus produced are amplified and transmitted to the distant end.

There they are amplified again and passed to a small stylus which rests lightly on a sheet of chemically prepared conducting paper ("Teledeltos") upon a revolving drum in exact synchronism with the drum at the sending point. As the stylus moves longitudinally along the revolving drum, the current passes through the paper with the result that the chemically treated surface of the paper is burned off leaving a black mark. Thus, exact reproductions of the transmitted matter are recorded and are at once available for use without the necessity of developing a film and making a photographic print.

A substantial portion of Western Union's business is derived from the private wire systems it designs, installs and leases to governmental and industrial customers. One of these, the AUTODIN system for the Department of Defense, is the world's most versatile and largest digital data communications network. AUTODIN is currently capable of transmitting 16 million punch cards daily, or the equivalent of 320 million words. Because it incorporates entirely new concepts and techniques in automated electronic switching, data can be exchanged from all kinds of transmission equipment—punch card machines, magnetic tape devices and teleprinters. This system is now being expanded to double its present size.

A smaller but somewhat similar system is the Advanced Record System for the General Services Administration scheduled for operation early in 1965. Initially, this computer-controlled network will handle $2\frac{1}{2}$ million messages yearly with a handling capability of 7 million annually.

E. F. SANGER

Cross-references: COMPUTERS, MICROWAVE TRANSMISSION, TELEMETRY.

TELEMETRY

The terms telemetry and telemetering imply both distance and measurement, but beyond this there is not universal agreement regarding the use and meaning of the terms. The American Standards Association defines telemetering as: "The measurement with the aid of intermediate means which permit the measurement to be interpreted at a distance from the primary detector." The distance involved may be anywhere

from a few inches in the case of certain test projectiles to many millions of miles in the case of space probes to other planets. Other terms such as data transmission system or data link are frequently used. A distinction in terminology is sometimes made between the transmission of a measurement for observation and interpretation and the transmission of a measurement to directly govern a controlling action. In modern applications, the quantity transmitted may not be a measurement at all, but the result of some complex calculation based on numerous measurements over a time interval.

The dominant transmission means for telemetry has been electrical, and a few applications were already reported during the first half of the nineteenth century. During the latter half of that century, applications included regular transmission of meteorological and other measurements. During the early part of the twentieth century, there were extensive applications of telemetry in connection with the Panama Canal followed by applications in many other areas such as electrical power and pipeline distribution systems. Soon after 1930, telemetry through the transmission medium of radio links was in use for meteorological measurements from small unmanned balloons. Soon after 1940, the first radio telemetering systems for testing aircraft were being designed and used. This occurred in response to the need for more complete and reliable measurement during experimental flights of high-performance military aircraft than could be provided by onboard recorders and pilot's observations. By 1945 radio telemetry was being used for measurements in the then small rockets. Subsequently, the development of advanced radio telemetering systems has been spurred by the well-known rapid development of the field of rocket-propelled vehicles. Recently, telemetry has seen rapidly increasing application in geophysical and biomedical research.

It is sometimes convenient to divide telemetry applications into two areas which might be called operational and test. By operational we mean the application of telemetry as a permanent part of some system in which case the telemeter must be compatible with the rest of the system in such aspects as cost, reliability, and maintainability. By test we mean the temporary application of telemetry to obtain test information during the developmental phase of some system in which case the telemeter need not be completely compatible with the rest of the system.

Telemetry systems for industrial purposes tend to fall more under the operational area and formed the majority of applications before 1940. Industrial telemetering systems have been frequently characterized by modest requirements for speed of response and accuracy which fall well within the physical limitations of the transmission link (frequently wire circuits) and are therefore accomplished by rather straightforward and uncomplicated signal handling methods. On the other hand, these applications have been characterized by demanding requirements for low first cost and

high reliability during long periods of unattended operation. Geophysical (including oceanographic) applications often have similar requirements. Industrial telemetry is well exemplified by the many thousands of telemeters in constant use by the electrical power industry throughout the world to measure electrical quantities and plant conditions.

The majority of aerospace telemetry during the last twenty years has been in the test area. Because of the destructive nature of tests and/or the difficulty in repeating them and because of the marginal nature of many of the systems being tested, aerospace telemetry has been characterized by demanding requirements for speed of response, accuracy, and number of channels which are frequently not well within the physical limitations of the radio transmission link. This together with severe weight and space restrictions and difficult environmental conditions for equipment operation has resulted in a field of telemetry activity more or less separate from the industrial area. Much attention has been given to the statistical efficiency of signal handling in the presence of random errors of various kinds and to the specialized hardware techniques needed to satisfy space, weight, power supply, and environmental conditions. The recent advent of long-lived satellites and space probes, including many systems whose only mission is telemetered measurements, has greatly emphasized telemetry reliability considerations.

The deeper understanding of the nature of measurement and information that has occurred during the last two decades by application of statistical methods has had its effect on the telemetering field. In the beginning, the design and classification of telemetering systems tended to be in terms of the physical quantity measured, the physical quantities used for transmission and interpretation, and the transducers for conversion between these physical quantities, i.e., in terms of the hardware. Terms such as current, voltage, frequency, ratio, phase, impulse, etc., were and are used to describe some systems. Multiplexing (the transmission of many measurements over the same channel) was either frequency division (assignment of a frequency band to each measurement, non-overlapping with those of the others) or time division (assignment of a periodic time interval to each measurement, non-overlapping with those of the others).

The statistical systems point of view in telemetry is concerned with *what* is done to the signal rather than *how* it is accomplished physically. From this point of view a telemetering system may do any or all of the following in logical sequence: (1) make a measurement; (2) abstract from the measurement in the form of a signal those characteristics which are needed for the eventual interpretation or decision; (3) store this signal; (4) code (modulate) this signal to give it greater immunity to the errors in the transmission link; (5) receive the signal; (6) store the signal; (7) decode (demodulate) the received signal to best preserve those characteristics which

are needed for interpretation and decision; (8) store the result. A measurement is said to have been made on a system if after the measurement, the uncertainty regarding the quantitative state of the system is less than the uncertainty existing before the measurement. Uncertainty in the sense of information theory has an exact quantitative measure in terms of the probability distributions involved and can be replaced by a cost index when a cost criterion is available. Similarly, a measurement is said to have been telemetered when there is a reduction in uncertainty or cost on the basis of the data produced by the telemetering system.

Current radio telemetering systems use many methods of modulation in various combinations. These include amplitude, phase, frequency, pulse amplitude, pulse position, pulse duration, pulse frequency, and pulse code. Multiplexing is sometimes accomplished by more general orthogonal methods than frequency- and time-division. With the advent of solid-state active circuit components and, most recently, of microelectronic techniques there has been a revolution in telemetry hardware. The practicability of complex circuits made possible by solid-state components, the use of digital computers for data handling, and developments in improved coding and demodulation methods have brought about considerable use of pulse-code modulation telemetering systems during the last decade.

LAWRENCE L. RAUCH

References

- Borden, P. A., and Mayo-Wells, W. J., "Telemetering Systems," New York, Reinhold Publishing Corp., 1959.
- Nichols, M. H., and Rauch, L. L., "Radio Telemetry," New York, John Wiley & Sons, 1956.
- Stiltz, H. L., Ed., "Aerospace Telemetry," Englewood Cliffs, N.J., Prentice-Hall, 1961.
- Balakrishnan, A. V., Ed., "Space Communications," New York, McGraw-Hill Book Company, 1963.
- Middleton, D., "Statistical Communication Theory," New York, McGraw-Hill Book Company, 1960.

Cross-reference: MEASUREMENT, PRINCIPLES OF.

TEMPERATURE AND THERMOMETRY

Temperature is the degree of hotness or coldness as measured on a definite scale; it is that property of an object which determines the direction of heat flow when the object is placed in contact with another. The concept serves as a measure of the average kinetic energy of the molecules of an object due to heat agitation. Human sensory organs can furnish an approximate indication of air temperature or of the temperature of an object. However, this indication can be strongly affected by factors such as moisture in the air or the heat conductivity of the object. Thermometry is the measurement of temperature and utilizes changes in some property of an object as an indicator. Changes in pressure, volume, length,

electrical resistance, and electrical potential have all been used in thermometers.

Temperature Scales. Many different scales have been used in measuring temperatures. The most common of them are Celsius or centigrade, Fahrenheit, Réaumur, Kelvin or centigrade absolute, and Rankine. Symbols used for the units on these scales are °C, °F, °R, °K, and °R, respectively. The Kelvin scale is noteworthy because it is essentially the true thermodynamic scale. The scales are all based on one primary reference point, which represents the temperature at which sharp change occurs. This temperature is assigned a number and serves as a starting point. A second reference point determines the size of the degree. All of these scales use the same primary reference point called the "ice-point". This is the temperature at which pure macroscopic ice crystals are at equilibrium with pure, liquid water under air saturated with moisture and at standard atmospheric pressure (1.01325 bar or 1.013 250 dynes/cm²). For the various scales this point is assigned numerical values as follows: 0°C, 32°F, 0°Réaumur, 491.7°Rankine, and 273.15°K. The value of 273.15°K is set so as to bring the scale as close as possible to the true thermodynamic scale. It is based on the use of 273.16°K for the triple point of water as an international standard (see STATES OF MATTER). The secondary point is that temperature at which the vapor pressure of pure water is equal to standard atmospheric pressure. The temperature interval between the primary and secondary points is divided into degrees, and the number of them between the two points is 100 Celsius degrees, 180 Fahrenheit degrees, 80 Réaumur degrees, 180 Rankine degrees, and 100 Kelvin degrees. When working above the secondary point or below the primary point, an attempt is made to keep the degrees the same size as those between the fixed points. The Fahrenheit and Rankine degrees are five-ninths as big as the Kelvin and Celsius or centigrade degrees.

Conversion from Scale to Scale. The conversion of a temperature reading from one scale to another requires remembering that the degrees are not always the same size and that the zero mark does not represent the same temperature on the different scales. Methods or formulas for the conversions can then be derived.

To convert a reading from degrees centigrade to degrees Fahrenheit, multiply by 9/5 and then add 32.

To convert a reading from Fahrenheit to centigrade, subtract 32 and multiply by 5/9.

To convert a reading from centigrade to Kelvin, add 273.15.

To convert a reading from Kelvin to centigrade, subtract 273.15.

To convert a reading from Fahrenheit to Rankine, add 459.7.

To convert a reading from Rankine to Fahrenheit, subtract 459.7.

To convert a reading from centigrade to Réaumur, multiply by 4/5.

To convert a reading from Réaumur to centigrade multiply by 5/4.

Similar formulas can be derived for conversions not shown above, or the calculation can be made in steps, using the conversions above to go from one scale to another until the desired reading is obtained. If a large number of readings must be converted, the use of a table such as may be found in handbooks is preferable both from the standpoint of avoiding mistakes and as a time saving device.

Reference Temperatures. By international agreement, a series of reference temperatures has been established to assure uniformity in measurements when using various types of precision thermometers. This is necessary because of lack of uniformity in changes in volume or length or electrical properties with respect to changes in temperature. The international temperature scale between -190 and +660°C is based on the resistance of a standard platinum resistance thermometer using the following formula for resistance at a temperature t :

Below 0°C:

$$R_t = R_0[1 + At + Bt^2 + C(t - 100)t^3]$$

Above 0°C:

$$R_t = R_0(1 + At + Bt^2)$$

where A , B and C are arbitrary constants. Starting at 660°C and up to the gold point, a platinum-platinum rhodium thermocouple is used, and above that, an optical pyrometer (see THERMOELECTRICITY and PYROMETRY, OPTICAL). Basic reference points include the boiling point of oxygen (-182.970°C), the triple point for water (+273.16°K), the boiling point of sulfur (+444.600°C), the melting point of silver (+960.8°C), and the melting point of gold (+1063.0°C).

Utilization of Temperature Scales. The centigrade or Celsius scale is very widely used for scientific studies, and most thermometers for exacting work are calibrated in this scale. It was invented in 1742 by Anders Celsius and, though the name centigrade is still widely accepted, the preferred name is Celsius. The Fahrenheit scale is used in every day life in the United States and England, and is the usual scale for engineering and medical practice in those countries. The Réaumur scale is in limited use in France, Germany, and Russia. The Rankine Scale is used in some engineering work, particularly where an absolute scale is needed for such things as the calculation of the theoretical efficiency of engines. The Kelvin scale was originated by Lord Kelvin and was based on a consideration of the second law of thermodynamics. This leads to a temperature scale with the zero point at the temperature at which all the thermal motion of the atoms stops. By using this as the zero point or absolute zero and another reference point to determine the size of the degrees, a scale can be defined. The Comité Consultative of the International Committee of Weights and Measures selected 273.16°K as the value for the triple point for water. This set the ice-point at 273.15°K.

Thermodynamic Scale. Thermodynamically speaking, the thermal efficiency E of an engine is equal to the work W derived from the engine divided by the heat supplied to the engine Q_2 . If Q_1 is the heat exhausted from the engine,

$$E = (W/Q_2) = (Q_2 - Q_1)/Q_2 = 1 - (Q_1/Q_2)$$

where W , Q_1 , and Q_2 are all in the same units. A Carnot engine is a theoretical one in which all the heat is supplied at a single high temperature and the heat output is rejected at a single temperature. The cycle consists of two adiabatics and two isothermals (see CARNOT CYCLES AND CARNOT ENGINES.) Here the ratio Q_1/Q_2 must depend only on the two temperatures and on nothing else. The Kelvin temperatures are then defined by the relation

$$\frac{Q_1}{Q_2} = \frac{T_1}{T_2}$$

where Q_1/Q_2 is the ratio of the heats rejected and absorbed, and T_1/T_2 is the ratio of the Kelvin temperatures of the reservoir and the source. If one starts with a given size for the degree, then the equation completely defines a thermodynamic temperature scale.

A series of Carnot engines can be postulated so that the first engine absorbs heat Q from a source, does work W , and rejects a smaller amount of heat at a lower temperature. The second engine absorbs all the heat rejected by the first one, does work and rejects a still smaller amount of heat which is absorbed by a third engine, etc. The temperature at which each successive engine rejects its heat becomes smaller and smaller, and in the limit this becomes zero so that an engine is reached which rejects no heat at a temperature which is absolute zero. A reservoir at absolute zero cannot have heat rejected to it by a Carnot engine operating between a higher temperature reservoir and the one at absolute zero. This can be used as the definition of absolute zero. Absolute zero is then such a temperature that a reservoir at that temperature cannot have heat rejected to it by a Carnot engine which uses a heat source at some higher temperature.

Thermometry. The measurement of temperature can be accomplished by many devices with varying degrees of accuracy or convenience. Among the instruments most frequently used are those based on (1) expansion of a solid, a liquid, or a gas; (2) change of pressure in a gas or vapor kept at a constant volume; (3) the thermoelectric effect; (4) changes in electrical resistance; and (5) changes in the character of thermal radiation.

The familiar mercury-in-glass thermometer utilizes the expansion of mercury which expands much more rapidly than does its glass container. For a wider temperature range, alcohol, usually colored for easier visibility, can be substituted. Maximum thermometers, such as the clinical thermometer, have a very narrow constriction just above the bulb so that if the temperature goes down, the mercury column breaks at the constriction leaving the column in the tube above the

constriction to indicate the highest temperature reached. A minimum thermometer can be built by including a small rider or index. As the liquid in the thermometer contracts, the rider is pulled towards the bulb, its end staying just at the surface of the liquid. If temperature goes up, the liquid flows past the index, leaving it to indicate the lowest temperature reached.

Gases are used as the working fluid in two types of thermometer. One holds the pressure constant, and the change in volume provides the temperature indication. The other is usually more convenient to use. In it the volume of the gas is held constant, and changes of pressure indicate temperature changes. Gases which are close to ideal gases are used, and corrections can be made for nonideal behavior so that the readings follow the absolute thermodynamic scale. This constant volume type has been adopted as the practical standard for the measurement of temperature.

A bimetallic strip can be made from small strips of two different metals welded or riveted together side by side. If the two have different coefficients of expansion, changes in temperature will cause the bimetallic strip to bend as one metal expands more than the other. This can be used to move a pointer over a scale. By using the motion of one end of the strip to open and close electrical contacts, a thermostat results for electrically controlling devices such as furnaces, ovens, or air conditioning units.

The thermoelectric effect can be used to measure temperature because the difference in potential created at a junction varies as a smooth function of temperature. The junction or sensor can be made very small and with low heat capacity so that such a thermocouple can be used for the precise indication of rapid changes in temperature.

Resistance thermometers require a wire or strip of metal and a circuit for determining changes in the resistance of the strip or wire as temperature changes.

An optical pyrometer measures the thermal radiation from a hot object over a narrow wavelength region and uses this radiation as a measure of temperature. Measurements are made by comparing brightness with that of a standard or by using photoelectric detectors. Such pyrometers can be used for temperatures far above those which can be measured by other methods (see PYROMETRY, OPTICAL). Radiation pyrometers measure the total thermal radiation and use this as an indication of temperature. They tend to be less accurate than optical pyrometers.

Each type of thermometer has its own advantages and disadvantages, and the choice of which to use must be made on the basis of the requirements of each particular application.

ROBERT M. BESANÇON

Reference

Herzfeld, Charles M., Ed., "Temperature: its Measurement and Control in Science and Industry," Vol. 3, Parts I and II (1962), Part III (1963), New York, Reinhold Publishing Corp.

Cross-references: CARNOT CYCLES AND CARNOT ENGINES; EXPANSION, THERMAL; GAS LAWS; HEAT; PYROMETRY, OPTICAL; RADIATION, THERMAL; STATES OF MATTER; THERMOELECTRICITY.

THEORETICAL PHYSICS

A division of physics into the two broad categories of theoretical and experimental physics is very common. Both, of course, share the general aim of physics to describe and correlate the results of past experiments, to predict the numerical results of experiments yet to be performed, and to develop concepts and methods which enable one to encompass many diverse and related phenomena into a single coherent description. The questions which one can ask of a physical theory are therefore of a "how" type, for example, "How long will it take a falling stone to reach the ground?" or "How many degrees would the temperature of the water in this pail drop if I were to leave it on this block of ice for one hour?" Questions of a "why" type such as, "Why does the stone fall toward the earth when I release it?" are presumed to be of the type which cannot legitimately be asked of physics. In general terms, the role of theoretical physics is the development of concepts which can be represented by mathematical symbols and then manipulated by using the wide variety of mathematical tools available; hence, the principal characteristic of theoretical physics is the importance of mathematics in its formulation and methods. The intermediate stages of a typical calculation generally cannot be checked at every step by a corresponding experiment. The value of a particular theory can therefore only be justified by how well the final results predicted by it agree with experiment. Occasionally it has been found that two quite different theories will agree equally well with experiment; then the choice between them has been made on another basis, usually by favoring the simpler one.

Theoretical physics as a career in itself became important only in this century. Pioneers in the development of physics such as Galileo and Newton did notable experiments in addition to the creation of useful mathematical methods. Much of the initial theoretical effort was devoted to MECHANICS and culminated in the development of Lagrange's and Hamilton's equations of motion. An important synthesis occurred later when it was found to be possible to write Maxwell's equations for electromagnetism in the same Hamiltonian form which had been devised for mechanics. Consequences of the rise of the atomic theory of matter were such subjects as KINETIC THEORY and STATISTICAL MECHANICS in which the bulk properties of materials are calculated as averages of mechanical and electrical characteristics of a large number of atomic and molecular particles. QUANTUM MECHANICS, as a method of dealing with the wave properties of matter, was developed during the first thirty years or so of this century and made it possible to calculate individual atomic properties which had become

measurable as a result of greatly improved experimental techniques. The older forms of mechanics, which had been so successful up to then, survived as a limiting case of the newer quantum mechanics. At about this same time, theoretical physics became a recognized specialty and the number of theoretical physicists has subsequently greatly increased. Present day theoretical physicists work both on the solution of specific problems of practical interest by known methods, and on the development of new theories appropriate to the description of the large amount of newly obtained experimental data. For convenience, theoretical physics can be divided into subject matter fields such as mechanics, electrodynamics, statistical mechanics, quantum mechanics, quantum field theory, and relativity. Another convenient subdivision is into macroscopic and microscopic theories. A macroscopic theory generally deals solely with relations among the measured properties of matter in bulk such as its heat capacity and electrical conductivity; thermodynamics is such a theory. In a microscopic theory, on the other hand, one tries to account for these large scale features in a more "fundamental" way by obtaining them from atomic properties. Related to this approach is the unsolved question of whether the fundamental numerical constants of nature, such as the electric charge and the mass of the electron, must be left entirely to experimental determination or may some day be calculable directly from a suitable general theory.

Experimental physics and theoretical physics are constantly influencing each other. Theory is strongly dependent upon experiment because the results of experiments provide the motivation to develop a new theory or to try to combine several existing theories into a more complete and general one. In turn, theory makes important contributions by suggesting new experiments which can serve to check various aspects of the theory or by predicting a previously unknown effect. Much of the effort of theoretical physicists is devoted to the extension and investigation of their theories in an attempt to predict such new effects. If these are confirmed by experiment, they increase one's confidence in the theory. Experimental results which differ from theoretical predictions show the necessity of improving the theory—sometimes the disagreement is so great that radical revisions are necessary. A few experiments have become known as "crucial" experiments because they have provided an unambiguous comparison between the basic foundations of an accepted theory and experiment. A famous example is the MICHELSON-MORLEY EXPERIMENT on the dependence of the speed of light upon the direction of motion of the earth in its orbit. The flat disagreement between the results of this experiment and the predictions of the then current theory led Einstein to make his searching reexaminations of the fundamental concepts of space and time which resulted in his development of the theory of relativity. It is common practice for experimenters to try to find empirical

formulas which will describe their results, and it is a constant challenge to theory to account for and to derive these formulas. Such formulas are common in SPECTROSCOPY, and their simplicity as compared to the complexity of the measurements inspired many efforts to calculate them; the first success of this kind was attained by Bohr in his quantum theory of the spectrum of hydrogen. Occasionally, use has been made of "thought experiments." As the name implies, these are not actually performed but are imagined and then analyzed in detail. For example, Heisenberg considered the use of a microscope as an instrument for measuring the position of a particle, paying particular attention to the effect of the interaction between the particle and the light which scatters the incident light into the microscope thus enabling the particle to be located. Considerations such as these were extensively used in the early development of quantum mechanics for the formulation of the "uncertainty principle" which is a statement of the inherent limitations on the accuracy of simultaneous measurements of related variables such as position and momentum.

Generally, theoretical physicists spend much of their time trying to solve specific problems. Often a problem arises from the need to fit the results of a given experiment into the established theory. Many problems are devised by the theoretical physicist himself, since experience has shown that progress in theory has rarely been accomplished by means of a single brilliant stroke of exceptional generalization, but rather by the slower process of solving bits at a time. It is only later that these various problems and their solutions can be combined into a more general problem in which the previous ones are now special cases of the new formulation. A particular situation may require new methods, and it is only by handling simplified cases at first that one can obtain enough experience and facility to enable the more complex case of ultimate interest to be finally solved. Accordingly, much of the training of physicists involves solving specific problems which are incorporated into textbooks. An important attribute of a competent theoretical physicist, which is much more profound than being able to solve an already existing problem, is the ability to formulate useful and significant problems, i.e., the ability to ask the right questions.

Virtually all theoretical problems involve a high degree of idealization. Quite often, the experimental situation involves very many variables and specific details of widely varying importance. The initial task of the theoretical physicist is to try to estimate which variables are of principal significance so that he can then decide which can be safely neglected in his analysis. For example, a volume of interest may often be regarded as so large that its boundaries can be assumed to be infinitely far away so that their effects are negligible. In essence, then, what is sought is a reasonable and tractable approximation that is a fairly good substitute for the

actual state of affairs. Arguments involving whatever properties of symmetry the system may have often play an important role in the approach to a problem. If, for example, it can be assumed that the distribution of mass in the earth depends only on the distance from the center and not on the particular direction in which one proceeds from the center, then the gravitational attraction of the earth will also not depend on direction but will involve only the distance from the center. Using this consequence of the geometrical symmetry, it is possible to simplify the problem by restricting one's attention only to those possible solutions which do not depend on any angle or direction. Although there exist a wide variety of formal methods of solution which have been devised through the years, one should also recognize how great the importance of experience has been as is shown by the many successful uses of intuition, hunches, and inspired guesses. Sometimes, one is somewhat inexplicably led to try a particular solution, which, upon test, turns out to be either the correct solution to the problem or so near to it that it is a relatively simple matter to adjust the trial solution to make it correct. At times, these trials are based on the knowledge that a uniqueness theorem exists, i.e., it has been shown previously that a particular equation has only one possible solution. One can then be confident that, no matter how devious the means may have been by which the solution was obtained, it is the only possible one and it is not necessary to spend one's time considering other possibilities. Many problems can be solved by using a general solution which has been found for a more general problem and then reducing this solution to the specific one needed at the moment by making it satisfy the appropriate conditions at the boundaries of regions concerned; these boundary conditions then suffice to determine specific numerical values of parameters which appear in the mathematical form of the general solution and which previously had to be left undetermined.

Somewhat related to the value of intuition is the extensive use which is made of analogies in the solution of certain problems. It is often quite surprising how many diverse subjects and topics can be described by exactly the same form of mathematical equations; this is an example of the unifying role of theoretical physics. From a practical point of view, methods and concepts which have proved useful in one field can then be transferred bodily over into another, without, in many cases, it being necessary to change the terminology. As an example, many problems in coupled vibrating mechanical systems have been successfully treated by using methods which were originally developed to cope with coupled oscillating electrical circuits. Similarly, the motion of electrons through the lattice of ions in a metal can be related to the propagation of waves along a chain of masses connected by springs.

From a purely mathematical point of view, the student of theoretical physics soon finds that very many problems cannot be solved exactly

in terms of simple or well-known mathematical functions. Although, in principle, these problems could all be solved by numerical methods, it is often of more interest and value to have a mathematically simple form for the solution. Consequently, the use of approximations is very extensive. This often consists of an expansion of a solution in a power series and of keeping only the first few terms since the others are of negligible magnitude. Sometimes only a knowledge of the order of magnitude of a quantity is sufficient, and this can be estimated quite well in spite of the impossibility of obtaining an exact solution, which in itself may be so complicated as to obscure the underlying features. In recent years, the development and availability of high-speed computers has been of great value for theory. The computers enable one to obtain numerical solutions of problems which previously were not solved because conventional numerical methods would simply take too long. Computer solutions have also proved useful in indicating the direction which should be taken by analytical methods and in suggesting appropriate types of approximations.

ROALD K. WANGSNES

References

The following list of selected references should enable an interested reader to obtain a more detailed picture of the scope and methods of theoretical physics. Many of the points mentioned above are discussed in more detail in these books and other examples of specific problems are described and analyzed.

- Einstein, A., and Infeld, L., "The Evolution of Physics," New York, Simon and Schuster, 1937.
 Lindsay, R. B., and Margenau, H., "Foundations of Physics," New York, John Wiley & Sons, Inc., 1936.
 Wangsness, R. K., "Introduction to Theoretical Physics: Classical Mechanics and Electrodynamics," New York, John Wiley & Sons, Inc., 1963.
 Wangsness, R. K., "Introductory Topics in Theoretical Physics: Relativity, Thermodynamics, Kinetic Theory, and Statistical Mechanics," New York, John Wiley & Sons, Inc., 1963.
 Bohm, D., "Quantum Theory," New York, Prentice-Hall, Inc., 1951.

Cross-references: FIELD THEORY; HEISENBERG UNCERTAINTY PRINCIPLE; KINETIC THEORY; MATHEMATICAL PHYSICS; MATRIX MECHANICS; MEASUREMENTS, PRINCIPLES OF; MECHANICS; QUANTUM THEORY; RELATIVITY; SPECTROSCOPY; STATISTICAL MECHANICS.

THERMIONICS

Thermionics is the science of the emission of electricity from solids induced by high temperature.

While thermionic emission was undoubtedly observed long ago in the discharge of electrified particles near heated solids and considerably studied in the last half of the nineteenth century,

not much progress in understanding it was made because fundamental concepts of the electric current were lacking. At the beginning of the present century the existence of the electron was established, and since then thermionics progressed rapidly both in theory and applications so that now it is basic to many large industries. As a measure of its growth, there are now in use probably two billion thermionic tubes in a host of different applications, and the basic theory of thermionics is largely worked out.

Among the early workers with currents from hot electrodes were Hittorf (1869-1883) and Goldstein (1885), both drawing quite large currents, and Elster and Geitel (1882-1889) who worked with very small currents, both positive and negative, in their research on the phenomenon. Edison (1883) in his work on the incandescent lamp discovered current emitted from the hot carbon filament and proposed a use for it in a patent granted to him. This emission became known as the "Edison Effect." None of these men, however, knew what they had, supposing that they dealt with ions such as occur in electrolysis or gas discharges. It was not until 1897-1899 that the work of J. J. Thomson showed that the negative carriers of cathode rays were a new species of particle with mass about 1700 times smaller than that of the hydrogen ion. This was the electron. Drude (1900) suggested that electrons rather than metallic ions are the carriers of current in metals, and Thomson proposed that they are also the negative charges emitted by hot metals. O. W. Richardson made a study on this basis (1901) and derived two forms of the "Richardson equation" relating the emitted current density i to the absolute temperature T of the metal and a property of the metal expressed by the letter b . The equations are

$$i = A_1 T^{1/2} \exp\left(-\frac{b_1}{T}\right) \quad (1)$$

and

$$i = A_2 T^2 \exp\left(-\frac{b_2}{T}\right) \quad (2)$$

A_1 and A_2 are arbitrary constants, $\exp(z)$ stands for the base of natural logarithms $e = 2.718$ raised to the exponent or power z , and b relates to the work required to remove an electron from the inside to the outside of the metal surface. The derivation of Eq. (1) was made using classical mechanics and the Maxwell-Boltzmann distribution of energies, later shown not to hold for free electrons within a metal. Equation (2) was based on the experimental fact that the electrons do not share in the specific heat of the metal. The formula even now retains essentially its original form [Eq. (2)], with expressions for the constants given in terms of known quantities. Richardson verified the form of his formula by careful experimental tests. Neither he nor subsequent workers could discriminate between the two forms, but the second is thought to be on a better theoretical basis. Richardson also introduced the term "thermionic."

W. Schottky (1919) and S. Dushman (1923) derived an expression for the constant A_2 in Eq. (2):

$$A = \frac{2T_1 k^2 m e}{h^3} = 60.2 \text{ amperes/cm}^2 \text{ deg}^2 \quad (3)$$

Here k is Boltzmann's constant, m is the mass and e the charge of the electron, and h is Planck's constant. Later derivations of the expression take into account the Sommerfeld (1928) theory of metallic conduction, the Pauli exclusion principle, and the spin of the electrons to yield a value of A of 120 amperes/cm² deg². The value of the exponent becomes

$$\frac{b}{T} = \frac{e\phi}{kT} + \frac{W_a - W_1}{kT} \quad (4)$$

Here ϕ is called the work function of the metal, usually expressed as a measured quantity in volts; W_a is the work of moving the electron out against the surface barrier, and W_1 is the energy the electron may have had inside the metal. There is in addition a factor introduced to account for the reflection r of electrons at the inner surface of the metal. The present form of the Richardson equation is then; according to Nordheim (1929).

$$i = A(1 - r) \exp\left(-\frac{e\phi}{kT}\right) \quad (5)$$

The exponential term of the expression has been amply verified over a large temperature range. The work function ϕ is known for a large number of metals. It tends to be larger for metals in which the atoms are packed closely together, smaller for open lattice metals, the range being about 1.5 to 6 volts. It is slightly higher on dense crystal faces than on open ones. The constant A may appear to vary by a large amount from its theoretical value of 120, in the range of 10 to 100 for pure metals. Actually, it is not that A is different but that ϕ varies with temperature. The pure metal that is most often used as a thermionic emitter where ruggedness and high voltage are involved is tungsten, because of its strength and high melting point. For it, the values of ϕ and apparent A are about 4.5 volt and 100 amperes/cm²deg². Molybdenum, tantalum and niobium are others used in special applications.

The thermionic properties of a metal surface are profoundly changed by thin films of foreign materials. This is the basis of the thoriated tungsten emitter used in small- and medium-size vacuum tubes by Langmuir and Rodgers in 1914. The filament is made of tungsten having a small additive of thorium. In the heat treatment, some of the thorium diffuses to the surface where it forms a quite stable deposit that is less than one atom deep. The work function of this surface is less than that of either thorium or tungsten, about 2.6 volts, and A is about 3 ampere/cm² deg². Where the tungsten filament is normally operated near 2700°K the thoriated tungsten yields a comparable emission current density

at 1800 to 2000°K, with a very considerable saving in heating power. The surface is less rugged than that of pure tungsten.

By far the larger number of vacuum tubes use the oxide-coated filament, described by Wehnelt (1904). On a metallic base that is usually a nickel alloy is deposited a relatively thick coating of the mixed oxides of barium, strontium and calcium. Certain activation processes yield a surface with work function ϕ in the region of 1 volt. In spite of the small and variable value of the factor A , in the range of .01, the low work function provides a surface of high emission so that it can be used at the temperature of near 1000°K, giving still higher thermal efficiency than the thoriated tungsten. The mechanism of electron emission from this surface is considerably different from that of the pure metal. The oxide layer is normally an insulator at room temperature that becomes a semi-conductor at the operating temperature. The oxides are partly dissociated so that there are metal atoms, particularly of barium, in the body of the layer and on its surface. The surface barium probably contributes to the low work function. The body barium is presumably ionized so that it contributes conduction electrons in the semi-conductor. At the metal-oxide interface, there is another low work function surface so that electrons can pass from the metal base into the oxide, to be available for emission at the outer surface. With this modified mechanism, the emission equation still essentially holds. The reason is largely that the Boltzmann factor $\exp\left(\frac{e\phi}{kT}\right)$

varies so rapidly with temperature that it renders other factors of little consequence experimentally.

So far it has been assumed that there is another electrode nearby with high enough positive voltage on it so that all of the emitted electrons are drawn to it. The current is then said to be saturated. In this condition, the current can still increase slowly with increasing voltage. The reason is that the strong electric field at the surface of the metal penetrates between the surface atoms and helps the electrons escape. The actual current then is increased above the saturation current by a factor $\exp 4.40 \left(\frac{F^{1/2}}{T}\right)$ determined

by Schottky in 1914, where F is the field strength in volts per centimeter, as verified with the refractory metals. With thoriated tungsten, the increase is more rapid than this, and with the oxide cathodes, the increase is so rapid that it is hard to say when saturation sets in.

At still higher surface fields, of the order 10 million volts/cm, another emission effect sets in, whereby the electrons are drawn out of even the cold metal. This is FIELD EMISSION, q.v.

At voltages below that required for saturation, the repulsion between the negatively charged electrons tends to limit the current. This is the space charge region, the condition in which most thermionic devices work. The current then is fairly insensitive to temperature and other conditions at the cathode so long as the anode voltage

is well below the saturation value. In this condition, the current can be controlled by grids and other means. The space charge limited current between a plane emitting cathode and an anode at the distance d from it and at voltage, V , each of area 1 cm^2 , is given fairly closely by the expression

$$i = 2.33 \times 10^{-6} V^{3/2} / d^2 \text{ amperes} \quad (6)$$

derived by Child (1911) and by Langmuir (1913). The equation is modified for a cylindrical structure, but the $V^{3/2}$ factor applies to any structure.

When the potential between emitter and plate is reversed so as to become retarding for electrons, the current is limited to the number of electrons with enough energy to overcome the retarding potential and is not limited by space charge. The current i_r is then related to the saturation current i_s by the expression

$$i_r = i_s \exp\left(-11600 \frac{V_r}{T}\right) \quad (7)$$

T being the temperature of the emitter and V_r the retarding potential. This may be written

$$\log i_r = -5030 \frac{V_r}{T} + \log i_s \quad (8)$$

Besides giving a means of determining the temperature of the emitter as shown by Germar (1925), the formula also is at the basis of electronic devices with logarithmic response.

The emission of positive ions from hot bodies seemed, before the existence of the electron was recognized, to be as important as the negative emission, and much work was devoted to it. It turned out eventually that it was not an important property of the body of the emitter but rather one of surface impurities. Richardson (1903-1914) showed that the positive ions given off by hot metals were ionized alkali atoms, and with a simple mass spectrometer he determined that they were mostly potassium coming originally from the glass envelope of the tube. Later studies have amply verified this finding, and the reason for it is now understood. A surface atom on a hot metal will be evaporated as a positive ion if the ionization potential of the atom is less than the work function of the metal, the metal then retaining the electron. This condition is satisfied with high work function metals as a base and with potassium, rubidium, cesium and possibly sodium as the impurity atoms.

There have been few practical applications of the positive ion emission beyond that of Kunsman (1927), but it has been useful in certain experimental researches such as those of Langmuir and Kingman (1925).

J. B. JOHNSON

References

- Richardson, O. W., "Emission of Electricity from Hot Bodies," New York, Longmans, Green & Co., 1916, 1921.
- Reimann, A. L., "Thermionic Emission," New York, John Wiley & Sons, Inc., 1934.
- Millman, J., and Seely, S., "Thermionics," New York, McGraw-Hill Book Co., 1941.
- de Boer, J. H., "Electron Emission and Adsorption Phenomena," New York, The Macmillan Co., 1935.
- Dushman, S., "Thermionic Emission," *Rev. Mod. Phys.*, **2**, 381 (1930).
- Becker, J. A., "Thermionic Electron Emission," *Rev. Mod. Phys.*, **7**, 95-128 (1935).
- Herring, C., and Nichols, M. H., "Thermionic Emission," *Rev. Mod. Phys.*, **21**, 185-270 (1949).

Cross-references: ELECTRON, ELECTRON TUBES, FIELD EMISSION, IONIZATION, PHOTOELECTRICITY.

THERMODYNAMICS

Classical Thermodynamics is a theory which on the basis of four main laws and some ancillary assumptions deals with general limitations exhibited by the behavior of macroscopic systems. Phenomenologically it takes no cognizance of the atomic constitution of matter. All *mechanical* concepts such as kinetic energy or work are presupposed. Thermodynamics is motivated by the existence of dissipative mechanical systems. A *thermodynamic system* K may be thought of as a collection of bodies in bulk; when its condition is found to be unchanging in time (on a reasonable time scale) it is *in equilibrium*. It is then characterized by the values of a finite set of say n physical quantities, it being supposed that none of these is redundant. Such a set of quantities constitutes the *coordinates* of K , denoted by $x (=x_1, \dots, x_n)$. Any set of values of these is a *state* \mathfrak{S} of K . In virtue of these definitions, K is in a state only when it is in equilibrium. The passage of K from a state \mathfrak{S} to a state \mathfrak{S}' is a *transition* of K . A transition is *quasi-static* if in the course of it goes through a continuous sequence of states, and if the forces which do work on the system are just those which hold it in equilibrium. A transition is *reversible* if there exists a second transition which restores the initial state, the final condition of the surroundings of K being the same as the initial condition. Reversible transitions are assumed to be quasi-static.

An enclosure which is such that the equilibrium of a system contained within it can only be disturbed by mechanical means is *adiabatic*, otherwise it is *diathermic*. For instance, stirring, or the passage of an electric current, constitute "mechanical means." A system K_0 in an adiabatic enclosure is *adiabatically isolated* but this does not preclude mechanical interactions with the surroundings. Its transitions are then called *adiabatic*.

For the time being, the masses of all substances present will be supposed fixed, and to achieve simplicity it will be given that (1) there are no substances present whose properties depend on their previous histories; (2) capillary forces as well as long-range interactions are absent. Further it will be supposed that of the n coordinates of K just $n - 1$ have geometrical character (*deformation coordinates*, e.g., volumes of enclosures), so

that the work done by K in a quasi-static transition is

$$\int dW = \int_{k=1}^{n-1} P_k(x) dx_k \quad (1)$$

Such a system will be called a *standard system* ($n-1$ enclosures in diathermic contact, each containing a simple fluid, may serve as example, x_n being any one of the pressures).

The Zeroth Law. Suppose two systems $K_A(x)$ and $K_B(y)$ to be in mutual diathermic contact. Experience shows that the states \mathfrak{E}_A and \mathfrak{E}_B cannot be assigned arbitrarily, but that there exists a necessary relation of the form

$$f(x; y) \equiv f(x_1, \dots, x_n; y_1, \dots, y_m) = 0 \quad (2)$$

between them. If K_C is a third system, its diathermic equilibrium with K_B on the one hand, or with K_A on the other, is governed by conditions

$$g(y; z) = 0 \quad (3)$$

and

$$h(z; x) = 0 \quad (4)$$

respectively. That these three functions are not independent is expressed by the *Zeroth Law*: *If each of two systems is in equilibrium with a third system then they are in equilibrium with each other.* It follows that any two of Eqs. (2) through (4) imply the third, i.e., they must be equivalent to equations of the form

$$\xi(x) = \eta(y) = \zeta(z) \quad (5)$$

Thus with each system there is now associated a function, its *empirical temperature function*, such that two systems can be in equilibrium if and only if their *empirical temperatures* (i.e., the values of their empirical temperature functions) are equal. Write $t = \xi(x)$; so that one has the *equation of state* of K_A . Also, t may be introduced in place of any one of the x_k . Note that the empirical temperature is not uniquely determined since $t_A = t_B$ may be replaced by $\phi(t_A) = \phi(t_B)$ where the function ϕ is monotonic but otherwise arbitrary: one has a choice of *temperature scales*. For a system not in equilibrium temperature is not defined.

The First Law. It is obvious that one can do mechanical work upon a system (say by stirring) while its initial and final states are the same. (Nothing is being said about the surroundings!) In this sense mechanical energy is not conserved. One might however hope that it is conserved at least in a restricted class of transitions. That this is so is asserted by the *First Law*: *The work W_0 done by a system K_0 in an adiabatic transition depends on the terminal states alone.* Thus if $\mathfrak{E}'(x')$, $\mathfrak{E}''(x'')$ are the terminal states

$$W_0 = F(x'; x'')$$

If $\mathfrak{E}'''(x''')$ is a third state, and the previous transition proceeds via \mathfrak{E}''' , W_0 must not depend on x''' , i.e.,

$$F(x'; x''') + F(x'''; x'') \equiv F(x'; x'')$$

It follows that there must exist a function $U(x)$, defined to within an arbitrary additive constant, such that

$$F(x'; x'') = U(x') - U(x'') (= -\Delta U \text{ say})$$

$U(x)$ is the *internal energy function* of K . (To make sure that U is in fact defined for all states, one assumes that *some* adiabatic transition always exists between any pair of given states.) The energy of a compound standard system is the sum of the energies of its constituent standard systems. Further, U must be a monotonic function of t , and it is convenient to choose the scale of t such that $\partial U / \partial t > 0$.

When the transition from \mathfrak{E}' to \mathfrak{E}'' is adiabatic, $W_0 + \Delta U$ vanishes by definition of U . If the transition is not adiabatic and W is the work done by K , the quantity

$$\Delta U + W (= Q, \text{ say}) \quad (6)$$

will in general fail to vanish. Q is then called the *heat absorbed* by K . Every element of a quasi-static adiabatic transition is subject to $dQ = 0$, i.e., by Eqs. (1) and (6), to the differential equation

$$\sum_{k=1}^{n-1} \left(P_k(x) + \frac{\partial U(x)}{\partial x_k} \right) dx_k + \frac{\partial U}{\partial t} dt = 0 \quad (7)$$

The Second Law. Experiment shows that if \mathfrak{E}' and \mathfrak{E}'' are arbitrarily prescribed states, then it may be that no adiabatic transition from \mathfrak{E}' to \mathfrak{E}'' exists. When this is the case one says that \mathfrak{E}'' is *inaccessible* from \mathfrak{E}' , but \mathfrak{E}' is then accessible from \mathfrak{E}'' , as has been already assumed. The states may of course happen to be mutually accessible. The existence of states adiabatically inaccessible from a given state is asserted precisely by the *Second Law*: *In every neighbourhood of any state \mathfrak{E} of an adiabatically isolated system there are states inaccessible from \mathfrak{E} .* (This formulation of the Second Law is known as the *Principle of Carathéodory*.) *A fortiori* this law applies to quasi-static transitions, i.e., those which satisfy Eq. (7). It asserts there are states \mathfrak{E}'' near \mathfrak{E}' such that no functions $x_k(t)$ exist which satisfy Eq. (7) and whose values when $t = t''$ are just x_k'' , ($k = 1, \dots, n-1$). It is merely a mathematical problem (the Theorem of Carathéodory) to prove that this is the case if and only if there exist functions $\lambda(x)$ and $s(x)$, ($x_n \equiv t$) such that the left-hand member is identically equal to λds , where ds is the total differential of s . Thus, the Second Law entails that

$$dQ = dU + dW = \lambda ds \quad (8)$$

(dQ is of course not a total differential). s is called the *empirical entropy function* of K . It is not uniquely determined, since it may be replaced by any monotonic function of s . If two standard systems K_A and K_B in diathermic contact make up a compound system K_C , $dQ_C = dQ_A + dQ_B$, i.e., because of Eq. (8),

$$\lambda_A ds_A + \lambda_B ds_B = \lambda_C ds_C$$

By including s_A, s_B and the common empirical temperature t among the coordinates of K_C , one infers that

$$\lambda_A = T(t)\theta_A(s_A), \lambda_B = T(t)\theta_B(s_B),$$

$$\lambda_C = T(t)\theta(s_A, s_B)$$

The common function $T(t)$ is called the *absolute temperature function*, while

$$S_A(s_A) = \int \theta_A(s_A) ds_A$$

is the *metrical entropy* of K_A . The "element of heat" dQ of any standard system thus splits up into the product of a universal function of the empirical entropy and the total differential $dS(x)$ of the metrical entropy function:

$$T dS = dU + dW' \quad (9)$$

By multiplying T by a constant and dividing S by the same constant, T can be arranged to be positive.

If one now chooses $x_n = S$ and recalls that the $x_k (k < n)$ are freely adjustable, the Second Law would be violated if S were also adjustable at will (by means of non-static adiabatic transitions.) Taking continuity requirements into account, it follows that S can either never decrease or never increase. The single example of the sudden expansion of a real gas shows that it can never decrease. One has the *Principle of Increase of Entropy: The entropy of an adiabatically isolated system can never decrease.*

The Third Law. It is known from experiment that for given values of the deformation coordinates, the energy function has a lower bound U_0 . The question arises whether the entropy S has an analogous property. It is found in practice that the specific heats $\partial U / \partial T$ of all substances appear to go to zero at least linearly with T as $T \rightarrow 0$. This ensures that the function S goes to a finite limit S_0 as $T \rightarrow 0$. Experiment shows however further that as $T \rightarrow 0$, the derivatives of S with respect to the deformation coordinates also go to zero. In contrast with U_0 , S_0 has therefore the remarkable property that it is independent of the deformation coordinates. One thus arrives at the *Third Law: The entropy of any given system attains the same finite least value for every state of least energy.* One immediate consequence of this is that the so-called *classical ideal gas* (the product of whose volume V and pressure P is proportional to T , and whose energy is a function of T only) cannot exist in nature. Further, no system can have its absolute temperature reduced to zero. The Third Law is therefore a statement about the properties of functions, not of systems, at $T = 0$.

The practical applications of the theory just outlined divide themselves into two broad classes: (1) those which are based on the existence and properties of the functions U and S and some others related to them—all "thermodynamic identities" being merely the integrability condition for the total differentials of these functions;

and (2) those which are based on the Principle of Increase of Entropy: the entropy of the actual state of an adiabatically enclosed system being greater than that of any neighbouring "virtual" state.

The most important of the auxiliary functions just mentioned are the *Helmholtz Function*:

$$F = U - TS \quad (10)$$

the *Gibbs Function*:

$$G = U - TS + \sum_{k=1}^{n-1} P_k x_k \quad (11)$$

the *Enthalpy*:

$$H = U + \sum_{k=1}^{n-1} P_k x_k \quad (12)$$

sometimes called *thermodynamic potentials*. Then, e.g.,

$$dF = -S dT - dW$$

F therefore contains all available quantitative information about K , since

$$S = -\frac{\partial F}{\partial T}, \text{ and } P_k = \frac{\partial F}{\partial x_k} \quad (13)$$

The same is true of G for instance, since

$$S = -\frac{\partial G}{\partial T}, \text{ and } x_k = \frac{\partial G}{\partial P_k}$$

F and G are naturally taken as functions of x_1, \dots, x_{n-1}, T and of P_1, \dots, P_{n-1}, T , respectively. At times one speaks of F as the "Helmholtz free energy" and of G as the "Gibbs free energy." In an *isothermal* reversible transition, the amount W of work done by a system is equal not to the decrease of its energy U but to the decrease $-\Delta F$ of its (Helmholtz) free energy. In the presence of internal sources of irreversibility

$$W < -\Delta F$$

In considering physicochemical equilibria, that is to say, if one is interested in the internal constitution of a system in equilibrium when changes of phase and chemical reactions are admitted, one introduces the *constitutive coordinates* n_i^a ; this being the number of moles of the i th constituent C_i in the a th phase. The definitions of Eqs. (10) through (12) remain unaltered, for the n_i^a do not enter into the description of the interaction of the system with its surroundings. Let an amount dn_i^a of C_i be introduced quasistatically into the a th phase of the system. The work done on K shall be $\mu_i^a dn_i$. The quantity μ_i^a so defined is the *chemical potential* of C_i in the a th phase. It is in general a function of all the coordinates of K . Then, identically,

$$dG = \sum_{k=1}^{n-1} x_k dP_k - S dT + \sum \sum \mu_i^a dn_i^a$$

Integrability conditions such as

$$\partial \mu_i^\alpha / \partial T = - \partial S / \partial n_i^\alpha$$

are applications of the first kind. On the other hand, the minimal property of G , derived from the maximal property of S , requires that

$$\sum_i \sum_\alpha \mu_i^\alpha dn_i^\alpha = 0$$

when all virtual states differ only in the values of the constitutive coordinates. If the system is chemically inert, the dn_i^α are subject only to the requirements of the conservation of matter. One then concludes that if there are c constituents and p phases, i.e., $n + pc$ coordinates in all, then the number f of these to which arbitrary values may be assigned is

$$f = c - p + n$$

This typical application of the second kind is the Gibbs PHASE RULE (for inert systems.) This rule is often stated merely for systems with only two external coordinates ($n = 2$, e.g., $x_1 = P$, $x_2 = T$). There must then be no internal partitions within the system, nor may it, for instance, contain magnetic substances in the presence of external magnetic fields.

The beauty and power of phenomenological thermodynamics lies just in the generality and paucity of its basic laws which hold independently of any assumptions concerning the microscopic structure of the systems which they govern. Its quantitative content is limited to conditions of equilibrium. Its conceptual framework is too narrow to permit the description of the temporal behavior of systems, except in as far as it makes it possible to decide which one of any pair of states of an adiabatically enclosed system must have been the earlier state.

Statistical thermodynamics seeks to remedy these deficiencies by making specific assumptions about the microscopic structure of the system K , and relating its macroscopic behaviour to that of its atomic constituents. K is then to be regarded as an *assembly* of a very large number of particles, which, on a non-quantal level, is a mechanical system with, say, N degrees of freedom. A *microstate* of K is a set of values of its N coordinates and its N conjugate momenta. It is out of the question to measure all these at a given time. One therefore constructs a *representative ensemble* \mathcal{E}_K of K , which is an abstract collection of a very large number of identical copies of K . At any time t , the members of \mathcal{E}_K will be in different microstates. Let the fractional number of members of the ensemble whose microstates lie in the range dp, dq about p, q be $\phi dp dq$. Then ϕ is the *probability-in-phase*, and with $d\Gamma = dp dq$

$$\int \phi d\Gamma = 1 \quad (14)$$

The reason for this terminology is implicit in the *Postulate*: *The probability that a given assembly K will, at time t , be in a microstate lying in the range $d\Gamma$ about p, q , is equal to the probability $\phi d\Gamma$ that*

the microstate of a member of \mathcal{E}_K , selected at random at time t , lies in the same range.

The mean value $\langle f \rangle$ of a dynamical quantity f is defined to be

$$\langle f \rangle = \int f \phi d\Gamma$$

If N is sufficiently large, fluctuations about the mean will usually be negligible.

When K is in equilibrium ϕ must be constant in time, and this will be the case if it is a function of the (time-independent) Hamiltonian H of K . Ensemble averages are now assumed to coincide with temporal averages. When, in particular, K is in diathermic equilibrium with its surroundings one can show that ϕ must have the form

$$\phi = \exp[(\Phi - H)/\theta] \quad (15)$$

where Φ and θ are independent of p, q . Then

$$\theta \langle \ln \phi \rangle = \Phi - \langle H \rangle \quad (16)$$

and, because of Eq. (14)

$$\int d \exp[(\Phi - H)/\theta] d\Gamma = 0 = \langle d[(\Phi - H)/\theta] \rangle$$

where d refers to a variation of the macroscopic coordinates of K . Using Eq. (16) and its variation, the relation

$$-\theta d \langle \ln \phi \rangle = d \langle H \rangle - \langle dH \rangle \quad (17)$$

follows. Now $\langle H \rangle$ ($= \bar{U}$, say) is the total energy of the assembly, while $\langle dH \rangle$ is the average of the change of the potential energy, i.e., the work $-dW$ done by the external forces on K . If one writes

$$S = -k \langle \ln \phi \rangle$$

where k is a constant, Eq. (17) becomes

$$k^{-1} \theta dS = d\bar{U} + dW$$

This is identical with the phenomenological relation of Eq. (9) if one formally identifies S with \bar{S} , \bar{U} with \bar{U} and θ with kT . In this way, contact with the phenomenological theory has been established, and the quantities characteristic of the one theory have been *correlated* with that of the other. With this correlation, or interpretation, Φ becomes F . However, because of Eqs. (14) and (15)

$$F = -kT \ln \int \exp(-H/kT) d\Gamma$$

so that if only H is known, the integral on the right (the *partition function*), and thus F , can be calculated. The equation of state of a real gas can thus in principle be obtained from a knowledge of the forces operating within the assembly. This illustrates how the additional information put into the theory yields a correspondingly greater output. Phenomenologically such an equation of state might be written as

$$PV = \sum_{n=1}^{\infty} B_n(T) V^{1-n}$$

but here each of the *virial coefficients* B_1, B_2, \dots must be measured separately.

If the quantum mechanical behavior of matter is taken into account, the fact that one cannot assign precise simultaneous values to canonically conjugate quantities must produce modifications of the details of the statistical theory. However, it is not necessary to consider these here.

H. A. BUCHDAHL

Reference

Buchdahl, H. A., "The Concepts of Classical Thermodynamics," Cambridge, England, The University Press (to be published in 1966).

Cross-references: ENTROPY, HEAT CAPACITY, PHASE RULE, PHYSICAL CHEMISTRY.

THERMOELECTRICITY

Thermoelectricity is the subject dealing with the interaction between temperature gradients and electrical potential differences in solid or liquid materials. In the absence of a magnetic field, there are three thermoelectric effects—the Seebeck, Peltier, and Thomson effects.

The Seebeck effect was discovered by T. J. Seebeck in 1822. Consider a circuit made up of two different materials as shown in Fig. 1. The

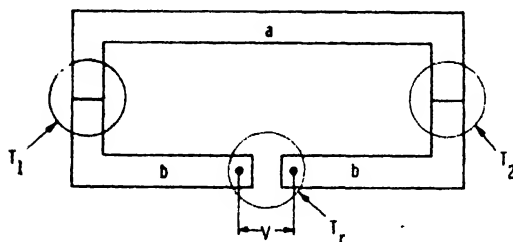


FIG. 1. A Thermoelectric circuit composed of two different materials, a and b. The regions enclosed in circles are assumed isothermal at the temperatures shown.

regions around the junctions are assumed isothermal at temperatures T_1 , T_2 , and T_R . If $T_1 \neq T_2$, a voltage V is observed. If the two materials a and b are homogeneous, the voltage V does not depend on T_R or on the temperature distribution along either material, but only on the temperature difference between the junctions. For small temperature differences, the voltage is proportional to the temperature difference

$$V_{ab} = \alpha_{ab}(T_1 - T_2) \quad (1)$$

where α_{ab} is called the Seebeck coefficient of the couple. (It has often been called the thermoelectric power, but this is a poor name since it does not have the dimensions of power.)

The Peltier effect was discovered by J. C. A. Peltier in 1834. In any conductor in which an electric current exists, heat is produced called the Joule heat which is given by

$$P_J = J^2 \rho \quad (2)$$

where P_J is the rate of Joule heat production per unit volume, J is the current density, and ρ is the electrical resistivity. At a junction between dissimilar materials, one finds an additional heat evolved or absorbed when current is present. This additional heat is called the Peltier heat and is given by

$$P_P = \pi_{ab} J_{ab} \quad (3)$$

where P_P is the rate of Peltier heat production per unit cross-sectional area of the junction and π_{ab} is the Peltier coefficient of the junction. Note that the Joule heat is always given off, whereas the Peltier heat may be absorbed or given off depending on the direction of the electric current.

The Thomson effect was predicted by William Thomson (later Lord Kelvin) in 1854 and experimentally established by him several years later. He found that a material in which there was a temperature gradient and an electrical current gave off or absorbed heat in addition to the Joule heat. The difference between the total heat given off and the Joule heat is called the Thomson heat. It is given by

$$P_t = \tau J \nabla T \quad (4)$$

where P_t is the rate of Thomson heat production per unit volume, τ is the Thomson coefficient of the material, and ∇T is the temperature gradient. Note that if either the temperature gradient or the electrical current is reversed in direction, then the Thomson heat reverses also, i.e., if Thomson heat originally was absorbed, then with a reversal of either the temperature gradient or the current density it will be emitted.

The three thermoelectric coefficients are related by the Kelvin relations

$$\pi_{ab} = T \alpha_{ab} \quad (5a)$$

$$\tau_a = -\tau_b = T(d\alpha_{ab}/dT) \quad (5b)$$

where T is the absolute temperature.

It should be noted that the Seebeck and Peltier coefficients as defined here involve two different materials, while the Thomson coefficient involves only one. There are two ways in which Seebeck coefficients for a single material are defined. (1) The relative Seebeck coefficient is defined as the Seebeck coefficient of a couple composed of the given material and a specified standard material such as platinum, lead, or copper. (2) The absolute Seebeck coefficient at temperature T_1 is defined by

$$\alpha_a(T_1) \equiv \int_0^{T_1} (\tau_a/T) dT \quad (6)$$

Since the Seebeck coefficient of a couple is zero at absolute zero temperature, integration of Eq. (5b) yields

$$\alpha_{ab} = \alpha_a - \alpha_b$$

At room temperature, metals have Seebeck coefficients in the range from a few tenths to as high as $40 \mu\text{V}/^\circ\text{C}$ for some alloys. Semi-metals such as bismuth have Seebeck coefficients ranging from about 20 to $40 \mu\text{V}/^\circ\text{C}$, while semiconductors

have Seebeck coefficients from a few microvolts per degree Celsius to as high as 1 mV/°C.

The Seebeck effect is widely used to measure temperature. The thermoelectric circuit, usually called a thermocouple, is made by welding together wires of pure metal or metallic alloys such as copper with constantan, chromel with alumel, or platinum with a platinum-rhodium alloy. For the measurements of very small temperature differences, it is possible to put a number of thermocouples in series so that the voltages add, as shown in Fig. 2. This device is called a thermopile.

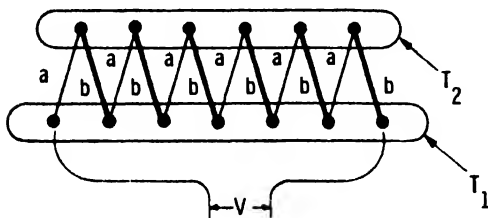


FIG. 2. A thermopile made from materials a and b. All of the upper junctions are at temperature T_2 and all of the lower ones are at temperature T_1 .

Devices have also been built utilizing the Seebeck effect to generate electricity directly from a heat source and utilizing the Peltier effect for refrigeration or heat pumping. In these applications the thermoelectric materials are semiconductors, such as Bi_2Te_3 , PbTe or GeTe , with a diameter of 1/8 to 1/2 inch and a length of 1/8 to 3/4 inch. The introduction of a magnetic field into a material may change its Seebeck, Peltier, and Thomson coefficients; it also produces several new effects, called galvanomagnetic and thermomagnetic effects. These include the Nernst, Ettingshausen, and Righi-Leduc effects. In these effects, the electric current or electrical potential difference, the magnetic field, and the temperature gradient or heat flow are all mutually perpendicular. These effects have also been used as a basis for devices which will pump heat or generate electricity directly from a temperature gradient (see HALL EFFECT AND RELATED PHENOMENA).

ROLAND W. URE, JR.

References

- Heikes, R. R., and Ure, R. W., Jr., "Thermoelectricity: Science and Engineering," New York, Interscience Publishers, 1961.
- MacDonald, D. K. C., "Thermoelectricity: An Introduction to the Principles," New York, John Wiley & Sons, 1962.
- Goldsmid, H. J., "Applications of Thermoelectricity," London, Methuen & Co., Ltd., 1960.
- Domenicali, C. A., *Rev. Mod. Phys.*, **26**, 237 (1954).
- Bridgman, P. W., "The Thermodynamics of Electrical Phenomena in Metals," New York, The Macmillan Company, 1934.
- Joffe, A. F., *Sci. Am.*, **199**, 31 (November 1958).

Angrist, S. W., *Sci. Am.*, **205**, 45 (December 1961).
 Wolfe, R., *Sci. Am.*, **210**, 70 (June 1964).

Cross-references: CONDUCTIVITY, ELECTRICAL; HALL EFFECT AND RELATED PHENOMENA; TEMPERATURE AND THERMOMETRY.

THIN FILMS

General. The term "thin films" is used for a wide variety of physical structures. Self-supporting solid sheets usually are called foils, when thinned from thicker material by such methods as rolling, beating, or etching, and films, when obtained by stripping a deposited layer from its substrate. Supported thin films are deposited on planar or (in special cases) curved substrates by such methods as vacuum evaporation, cathode sputtering, electroplating, electroless plating, spraying, and various chemical surface reactions in a controlled atmosphere or electrolyte. Thicknesses of such supported films range from less than an atomic monolayer to a few microns ($1\mu = 10^{-4}$ cm). A frequently used thickness measure is the angstrom ($1\text{\AA} = 10^{-8}$ cm). Thin films not forming a continuous sheet are called "island films." Particularly noble metals may condense as islands of considerable thickness (up to $\sim 10^2\text{\AA}$).

In scientific studies and technical applications, the use of well-controllable deposition methods such as vacuum evaporation and cathode sputtering are generally preferred. The film structure is markedly influenced by such deposition parameters as substrate composition and surface structure, source and substrate temperatures, deposition rate, and composition and pressure of the ambient atmosphere (where applicable). In general, the structure of films is more disordered than the corresponding bulk material. Smaller grains, higher dislocation concentrations, and deviations from stoichiometry are typical, and films approach bulk structure only as a limiting case. Under certain growth conditions, films exhibit preferential crystal orientations or even epitaxy. (Epitaxy means that the film structure is determined by the crystal structure and orientation of the underlying substrate.)

Solid thin films are common study objects in most phases of solid-state physics. They supply the samples for the study of general structural and physical properties of solid matter where special beam methods require small quantities of material or extremely thin layers, as for instance in transmission electron microscopy and diffraction, NEUTRON DIFFRACTION, UV spectroscopy, and X-RAY DIFFRACTION AND SPECTROSCOPY. Thin films represent the best means for studying physical effects, where these effects are caused by the extreme thinness of the material itself. Examples are the rotational switching of ferromagnetic films, electron tunnelling phenomena, electromagnetic skin effects of various kinds, and certain optical interference phenomena (see FERROMAGNETISM, SKIN EFFECT AND TUNNELING). Films also are convenient vehicles for the investigation of nucleation and crystal growth, and for

states of extremely disturbed thermodynamic equilibrium.

Presently, films find three major industrial uses: the decorative finishing of plastics, optical coatings of various kinds (mainly antireflection coatings, reflection increasing films, multilayer interference filters, and fluorescent coatings), and in electronic components from transistors over resistor-capacitor networks to such specialized devices as magnetic storage bits, photosensors, and cryotrons. The restricted space only permits the discussion of a few selected research and application areas.

Nucleation, Growth and Mechanical Properties of Films. In vacuum evaporation, molecules or atoms of thermal energy are deposited at a uniform angle of incidence and under well-defined environmental conditions. Most nucleation and growth studies, therefore, have been made on evaporated films. A particle approaching the substrate enters close to its surface a field of attracting short-range London forces with an exchange energy proportional to $-1/r^6$. At a still shorter distance r , repulsive forces proportional to $e^{-r/\text{constant}}$ resist the penetration of the electron clouds of the surface atoms. Due to the atomic or crystalline structure of the substrate, this potential field exhibits periodicity or quasi-periodicity in the substrate plane. The freshly condensed particles migrate over the surface with a jump frequency $i_D \propto \exp(-Q_D/kT)$, or desorb with a frequency $i_{ad} \propto \exp(-Q_{ad}/kT)$, where the activation energy Q_D is often approximately one-fourth of Q_{ad} . Permanent condensation occurs in most cases at distinct nucleation centres which may consist of deep potential wells of the substrate, clusters of condensed particles, or previously deposited "seed" particles of a different material. The number of nuclei formed in the second case is strongly temperature and rate dependent.

Most metals always condense in crystalline form, but the grain size is extremely small at low temperatures (in the order of a few angstroms) and increases markedly with increasing substrate temperatures. Grain size decreases with increasing deposition rates. The condensation of amorphous or quasi-liquid phases at low temperatures has been observed for such metals as antimony and bismuth and a few dielectrics. Some of these materials, on annealing, pass through otherwise unobserved and probably metastable phases.

Stresses of considerable magnitude are often observed in deposited films. The main causes of these stresses are a mismatch of expansion coefficients between substrate and film, enclosed impurity atoms, a high concentration of lattice defects and in very thin films, a variety of surface effects. Often, the stresses resulting from lattice defects can be minimized by the choice of a higher substrate temperature during deposition, or they can be reduced by a post-deposition anneal. Metal films frequently exhibit tensile strengths which are considerably larger than those of the corresponding bulk materials.

Thin-film Optics. Deposited metal mirrors probably represent the oldest optical application of

films. High-quality mirrors usually are produced by the vacuum evaporation of aluminium on an appropriately shaped glass substrate. Often, a glow-discharge cleaning of the substrate or a chromium undercoat is first applied to increase the adhesion of the aluminium. After deposition, the aluminium is protected by anodic oxidation or an evaporated overcoat of SiO, SiO₂, or Al₂O₃.

For SiO, maximum reflectance in the visible spectral region is achieved at a thickness of about 1400 Å. Rapid SiO evaporation reduces the reflectance at shorter wavelengths.

Single or multilayer coatings find increasing use as optical interference filters. These film stacks may consist solely of transparent films of different refractive indices n_i , or a combination of absorbing and nonabsorbing layers. Common low-index materials for glass coatings in the visible region of the spectrum are MgF₂ ($n_i = 1.32$ to 1.37), and cryolite Na₃AlF₆ ($n_i = 1.28$ to 1.34); high-index materials are SiO ($n_i = 1.97$), ZnS ($n_i \approx 2.34$), TiO₂ ($n_i = 2.66$ to 2.69) and CeO₂ ($n_i = 2.2$ to 2.4). The indices are given for the sodium D-line. Various semiconductors are used for infrared coatings.

At each air-film, film-film, or film-substrate interface, the incident light amplitude is split into a reflected and a transmitted fraction according to the Fresnel coefficients

$$f_{j-1} = (\hat{\mathcal{N}}_{j-1} - \hat{\mathcal{N}}_j) / (\hat{\mathcal{N}}_{j-1} + \hat{\mathcal{N}}_j) \text{ and } g_{j-1} = 2 \cdot \hat{\mathcal{N}}_{j-1} / (\hat{\mathcal{N}}_{j-1} + \hat{\mathcal{N}}_j)$$

where j and $j-1$ denote the number of the optical layer counted from the side of the incident beam. $\hat{\mathcal{N}}_j = \mathcal{N} / \cos \theta_j$ for p polarization or $\hat{\mathcal{N}}_j = \mathcal{N}_j \cdot \cos \theta_j$ for s polarization is the effective refractive index, and $\mathcal{N}_j = n_j - ik_j$ the refractive index of the j layer.

$$\cos \psi_j = \sqrt{(n_j^2 + k_j^2 + p_j^2) / 2} - i \sqrt{(n_j^2 + k_j^2 - p_j^2) / 2}$$

$$p_j = 1 + (k_j^2 - n_j^2) [n_0 \sin \theta_0 / (n_j^2 + k_j^2)]^2$$

$$q_j = -2n_j k_j [n_0 \sin \theta_0 / (n_j^2 + k_j^2)]^2$$

The symbol θ_0 is the angle of incidence in the incident medium.

For nonabsorbing film stacks ($k_i = 0$; $i = 1, 2, \dots, m+1$), the over-all reflectance and transmittance may be obtained by summing the multiple coherent reflections between the film boundaries. A more general treatment based on electromagnetic theory yields for amplitude reflectance and transmittance the recursions formulas

$$r_{(j-1)-} = (f_{j-1} + r_{j-} \exp(-2i \Phi_j)) / (1 + f_{j-1} r_{j-} \exp(-2i \Phi_j))$$

and

$$t_{(j-1)-} = (g_{j-1} t_{j-} \exp(-i \Phi_j)) / (1 + f_{j-1} r_{j-} \exp(-2i \Phi_j))$$

$\Phi_j = \Phi_j \cos \Theta_j$ is the effective phase thickness. $\Phi_j = (2\pi/\lambda) \mathcal{N}_j l_j$ where λ is the wavelength in vacuo and l_j is the geometrical film thickness. The recursion is started on the side of emergence, using the initial conditions $r_{m-} = f_m$ and $t_m = g_m$. Intensities are given by $R = |r_0|^2$ and $T = (\mathcal{R} \mathcal{N}_{m+1}/n_0) |t_0|^2$ where \mathcal{R} denotes "real part of." If A_j is the absorption in the layer j , $R + T + \sum_j A_j = 1$.

A single antireflection coating of $\lambda/4$ optical thickness $n_2 l_2$ yields zero reflectance at $n_2 = \sqrt{n_{\text{glass}}}$. A double layer coating of $\lambda/4$ films requires $n_2/n_1 = \sqrt{n_g}$. The transmission of a Fabry-Perot interference filter consisting of a dielectric spacer layer between two partially reflecting metal films is given by $I/I_0 = [(1 + A/T)^2 + (4R/T^2) \sin^2(\delta - \Phi)]^{-1}$ where $\Phi = 2\pi n l \cos \theta/\lambda$. R , T , and A are the reflection, transmission and absorption coefficients of the reflecting layers. The refractive index and thickness of the spacer film are n and l . θ is the angle of refraction in the spacer, and δ the phase change for reflections at the spacer-metal film interfaces. $(I/I_0)_{\text{max}} = (T/(1 - R))^2$ and $(I/I_0)_{\text{min}} = (T/(1 + R))^2$. The band pass half-width is $\Delta\lambda_{1/2} \approx \lambda (1 - R)/m\pi R^{1/2}$ for the interference order m ($m\pi = \Phi$). More, complex coatings and filters, and their various applications, cannot be discussed here. It should be mentioned, however, that films play a very important role today in the accurate determination of the optical constants of many materials, but particularly of metals (see REFLECTION).

Film Electronics. Deposited dielectric film materials in common use are SiO_2 , MgF_2 , ZnS , and various organic compounds. Thin capacitive layers in the 100 to 500 Å thickness region are often produced by the anodization of tantalum and aluminum to Ta_2O_5 or Al_2O_3 , respectively. The breakdown strength and dielectric constant of films approach bulk values, but might be reduced by surface roughness, structural faults, and lower density. According to the Lorentz-Lorenz formula, the dielectric constant D changes with reduced density ρ as $dD/d\rho = 3C/(1 - C\rho)^2$, where C is a constant depending on the material. On metal-dielectric-metal films, quantum mechanical tunneling through the dielectric film becomes observable below a dielectric thickness of about 100 Å. For applied voltages less than the metal-insulator work function ϕ , the tunneling current density J is proportional to the applied voltage V , demonstrating that the low-voltage tunneling resistance is ohmic. $J = (qV/h^2 s)(2m^* \phi)^{1/2} \exp[-(4\pi s/h)(2m^* \phi)^{1/2}]$. At high applied voltages ($qV > \phi$), the current increases very rapidly: $J = (q^2 V^2/8\pi h \phi_s^2) \exp[-(8\pi s/3h q V)(2m^*)^{1/2} \phi^{3/2}]$. s is the insulator thickness, m^* the electronic effective mass, and q the electron charge. Thicker dielectric films may exhibit in high fields appreciable Schottky or avalanche currents when they are greatly disordered.

Polycrystalline metal films generally show, due to their low structural order, a larger resistivity than the bulk material. According to Matthiessen's rule, the total resistivity can be expressed as

$\rho = \rho(t) + \rho(i)$ where $\rho(t)$ is the temperature-dependent resistivity associated with scattering by lattice vibrations, and $\rho(i)$ is a temperature-independent resistivity caused by impurity or imperfection scattering. Very thin specimens with a thickness comparable to the electron mean free path show a $\rho(i)$ rapidly increasing with decreasing thickness. This increase is caused by an increasing contribution of non-specular electron scattering at the film surfaces. By annealing a metal film, $\rho(i)$ might be reduced permanently. A large $\rho(i)$ results in a small temperature coefficient α .

Many known superconductors can be deposited as super-conductive films (see SUPERCONDUCTIVITY). Through thin-film experiments, the energy gap in semiconductors can be measured, and material parameters, such as the penetration depth of magnetic fields, can be studied at dimensions less than the coherence range.

Studies of semiconductor films have shown many facets. The properties of epitaxial films have mainly been investigated on Ge and Si, and to a lesser degree on III-V compounds. Much work has been done on polycrystalline II-VI films, particularly with regard to the stoichiometry of the deposits, doping and post-deposition treatments, conductivity and carrier mobility, photo-conductance, fluorescence, electroluminescence, and metal-semiconductor junction properties. Among other semiconductors, selenium, tellurium, and a few transition metal oxides have found some interest.

Film resistors, capacitors, and interconnected R-C networks on planar glass or ceramic substrates are finding widespread industrial use. Common resistor materials are carbon, nichrome and tin oxide in individual components; and nichrome, tantalum, tantalum nitride, SiO-chromium cermet, and cermet glazes in planar networks. Gold, copper, aluminum, or tantalum are used for termination lands, connection leads, and capacitor plates. SiO , MgF_2 and Ta_2O_5 serve as film capacitor dielectrics and crossover insulation. The geometrical configuration of the desired component or circuit pattern is obtained either by deposition through mechanical masks or by removing from a continuous sheet the undesired portions after the deposition process is completed. This removal is frequently accomplished by a combination of photolithographic and etch processes.

The minimum length l and width w of a resistor are calculated from the given resistance R , the sheet resistance \mathcal{R} in ohms per square, dissipated power P , and permissible power dissipation per square inch \mathcal{P} by use of the formulas $w = \sqrt{(P \cdot \mathcal{R})/\mathcal{P} \cdot R}$ and $l = wR/\mathcal{R}$. The capacitance of film capacitors is given by $C = 0.225 D(N-1)A/t$, where C is the capacitance in picofarads, D the dielectric constant, N the number of plates, A the area in square inches, and t the dielectric thickness in inches.

Thin-film semiconductor devices have not yet reached the production stage, mainly due to difficulties in controlling the film surface and

interface properties. Various barrier layer diodes have exhibited impressive rectification ratios, but limited breakdown strength and low speed due to their large specific capacitance. Of the many film TRANSISTOR concepts proposed, the insulated gate field effect device looks the most promising and manufacturable. Its structure consists of a minute metal-dielectric-semiconductor capacitor. The semiconductor strip carries current between two terminals called source and drain. A field applied between metal "gate" and source modulates the semiconductor conductance and consequently the source-drain current. Usable semiconductor materials with a sufficiently low concentration of interface states are CdS, CdSe, and tellurium. These devices exhibit pentode-like characteristics with voltage gains ranging from 2.5 at 60 Mc to 8.5 at 2.5 Mc. The gain band width product G.B., which is equal to the transconductance divided by 2π times the gate capacitance, reaches values of about 20 Mc. It is determined by $G.B. = \mu_d V_D / 2\pi L^2$, where μ_d is the effective drift mobility of the electrons, V_D the source-drain potential, and L the source-drain spacing which is usually chosen between 5 and 50μ . Special film semiconductor devices in industrial use are various types of photodetectors.

Magnetic Films. Magnetic thin films of nickel-iron (usually desposited at an 80:20 composition by weight) exhibit a number of unusual properties, which have led to many experimental and theoretical studies, as well as to important applications in binary storage and switching, magnetic amplifiers, and magneto-optical Kerr-effect displays.

Such "Permalloy" films have two stable states of magnetization, corresponding to positive and negative remanence. When deposited in a magnetic field or at an oblique angle, they exhibit uniaxial anisotropy. In practice, this anisotropy shows some dispersion, since it results from the alignment of local lattice disturbances. The stable states result from the minimization of the free energy $E = MH_L \cos \theta - MH_T \sin \theta - K \sin^2 \theta$, where the last term represents the anisotropy energy, and θ is the angle between the magnetization M and the easy axis. From an inspection of the derivatives of this equation follows the hard-direction straight-line and the easy-direction square hysteresis loops of anisotropic films. In the latter case, the magnetization is always either $+M$ or $-M$, and the change occurs at $H_L = \pm H_K$. The transitions from unstable to stable states occur at $\partial^2 E / \partial \theta^2 = 0$, resulting in a critical curve $H_L^{2/3} + H_T^{2/3} = H_K^{2/3}$ which has the form of an asteroid enclosing the origin (see MAGNETISM).

An important feature of magnetic films is the high speed with which the state of magnetization can be reversed. Dependent on film properties and magnetic fields, three modes of magnetization reversal occur: Domain wall motion, incoherent rotations, and the extremely fast coherent rotation of the magnetization. Wall-motion switching is expected when the driving fields are smaller than the critical values. Modern high-

speed film computer memories attempt to utilize the rotational switching mode.

RUDOLF E. THUN

References

- Dushman, S., "Scientific Foundations of Vacuum Technique," Second edition, New York, John Wiley & Sons, 1962.
- Holland, L., "Vacuum Deposition of Thin Films," New York, John Wiley & Sons, 1958.
- Keonjian, Edward, Ed., "Microelectronics," New York, McGraw-Hill, 1963.
- Hass, G. Ed., "Physics of Thin Films," Vols. I and II, New York, Academic Press, 1963 and 1964.
- Neugebauer, C. A., *et al.*, Ed., "Structure and Properties of Thin Films," New York, John Wiley & Sons, 1959.
- Mayer, H., "Physik dünner Schichten, I and II," Stuttgart, Wissenschaftliche, 1950 and 1955.
- Heavens, O. S., "Optical Properties of Thin Solid Films," London, Butterworths, 1955.
- Series: "Vacuum Technology Transactions," New York, Pergamon Press, 1955-1963.
- Walter, H., in Flüge, S. Ed., "Handbuch der Physik," Vol. 24, Berlin, Springer, 1956

Cross-references: CONDUCTIVITY; ELECTRICAL; ELECTRON MICROSCOPE; FERROMAGNETISM; MAGNETISM; NEUTRON DIFFRACTION; SEMICONDUCTORS; SKIN EFFECT; SPECTROSCOPY; SUPERCONDUCTIVITY; TRANSISTOR; TUNNELING; X-RAY DIFFRACTION.

TIME

Too many interpretations of the concept of time are based on one of the following two kinds of oversimplification. Philosophers have speculated about time on the premise that it is a primary notion and can be abstractly defined without bothering about the implementation of the definition. Conversely, the physicists, before Einstein, had a tendency to take time for granted and to use it as a parameter without further questioning its definition. Psychologists may have been the first to make a step in the right direction by trying to relate the concept of time to actual perceptions.

Nowadays, the physicist has become aware of the necessity of providing operational definitions of the concepts he uses, and it is generally acknowledged that the very concept of time depends upon the possibility of the repetition of events that may be considered identical or, at least, that have a common recognizable feature. However, the far-reaching implications of this idea are seldom realized, and it is not infrequent to read otherwise respectable discussions based on a notion so "obvious" that its vagueness remains completely unsuspected, namely, the notion of a "clock."

Assigning to time the character of a self-contained concept and assuming the existence of appropriate "clocks" showing the flow of this "time" is putting the cart before the horse. In a more refined approach, an a priori time

concept is accepted and principles are formulated by virtue of which a motion taking place in some specified conditions is uniform. But even such a procedure amounts to a self-deception, as the actual definition of time is then camouflaged behind those "principles." (For example, the essential of the definition of time in classical mechanics lies in Newton's first law). In brief, the true primary operation is the *arbitrary choice* of a repeatable phenomenon that may be used in the definition of a clock. The rate of flow of "time" is then implied by this choice, that is, the choice of the fundamental clock *is* the definition of time.

For practical purposes, it is appropriate to limit the freedom inherent in the choice of a clock by specifying convenient properties to be imposed on the resulting time scale. The main such properties are the availability of a sufficiently perennial master "clock" and the possibility of devising wieldy secondary clocks, for everyday use, that give reproducible and consistent readings endowed with a property of additivity. The first master clock that suggested itself to mankind was provided by the rotation of the earth on its axis and around the sun. The unit of time thus defined and its aliquot parts gave birth to the first astronomical time scales (mean solar day, tropical year), which served as a background for the development of classical mechanics and astronomy. A large number of phenomena were discovered (e.g., the beats of a good watch) that bear a linear relationship to that time scale. Any of the "linear systems" involved could be used as a secondary clock. Then, when the measuring techniques improved, it appeared that the mutual linearity of those phenomena was only an approximation. This discovery brought about a mild crisis of the metrology of time. The crisis was readily dismissed by stating that the rotation of the earth was not really uniform (as compared to "more accurate" clocks). In fact, the situation had a deeper purport: two descriptions of the fundamental master clock, that had been hitherto considered equivalent, appeared to be inconsistent, and the question of the *choice* of the clock was brought to the foreground again. The difficulty was temporarily settled in 1955 by relating the astronomical time scale to one particular period (tropical year 1900.0 \approx 31 556 925.975 seconds), and work is under way in order to submit a new definition of the second to the Twelfth International Conference of Weights and Measures in 1966. This definition will be based on the frequency of the radiation emitted in the transition between two specified atomic energy levels. Implemented by a so-called atomic clock, it will yield standards reproducible up to a relative error of 2×10^{-11} .

Newtonian time, a basic feature of the whole body of classical mechanics, is practically the astronomical time, operationally defined as above, complemented by the following extra postulate. Let any single observer, standing still on earth, determine the (improved as above)

astronomical time scale; the time scale so obtained is then to be used by every "observer," whatever his motion with respect to the first one. This postulate expresses, for each "observer," one choice of the master clock among infinitely many possible choices, and as such it is legitimate. When it was stated, it was also consistent with the contemporaneous physical knowledge. Later on, the physicists grew accustomed to certain properties derived from the choice of Newtonian time and space and from other postulates of mechanics and electromagnetism, until a calamity happened which was similar to what has befallen astronomical time: the improvement of the measuring techniques showed that one empirical fact (namely, that the velocity of light in the laboratory frame of reference is independent of the motion of the emitter) was not compatible with all of those properties. Again a choice was necessary. The analysis of special relativity disclosed which of the properties at stake were incompatible. The decision as to which to drop was largely a matter of convenience. Einstein's choice (justified by strong operational reasons) was to drop the universality of time and space in order to retain more physical postulates. From then on, time and space ceased to be absolute concepts (see RELATIVITY). It is worth mentioning that the presence of such a fundamental choice at the basis of the special-relativistic theory of time is not always recognized.¹

As the choice of the master clock may not be made any more by one "observer" on behalf of another one, the choice has to be decided upon for each "observer" separately. Special relativity's specification is that each inertial "observer" shall use the time scale defined by means of a conventional atomic "clock" at rest with respect to himself. This procedure leads to the concept of proper time, which enjoys a mathematically invariant character and plays, for inertial systems, the part formerly played by the universal time (see RELATIVITY).

When it comes to comparing the descriptions of the universe as made by "observers" whose motions relative to an inertial frame involve an acceleration, use is generally made of "general covariance," which permits a straightforward generalization of proper time. The relevant mathematical framework is that of general RELATIVITY. However, the interpretation of the general-relativistic proper time in terms of everyday experience involves instantaneous inertial frames, which are not embodied by any material system in this kind of problem. Therefore, the explicit relationship between proper time and the time actually measured by a material device called a clock is by no means clear.² In fact, the use of proper time in accelerated systems again implies a choice of the master clock, and in the present instance the choice has the drawback of being fairly abstract. Notwithstanding the largely widespread opinion, general covariance may not be the most convenient tool for the study of time in noninertial frames of reference. The problem of time measurements in

such frames is an open problem, on which practically everything remains to be done.

Although quantum mechanics has not brought a revolution in the conception of time (which remains an evolution parameter in the physical equations), it has suggested one approach which, if it ever turns out to be fruitful, would seriously upset the current notion of time. This approach consists in an attempt to quantize space-time itself, i.e., to assign to space and time jointly a discontinuous structure, involving elementary lengths and durations, instead of simply quantizing phenomena in a continuous space-time framework. As no consistent formalism of practical use has been developed so far along these lines, it suffices to mention a couple of recent references.^{3,4}

JACQUES E. ROMAIN

References

1. Romain, J. E., *Nuovo Cimento*, **30**, 1254 (1963).
2. Romain, J. E., *Rev. Mod. Phys.*, **35**, 376 (1963); *Advances in Astronautical Sciences*, **13**, 616 (1963); *Nuovo Cimento*, **31**, 1060 (1964); **33**, 1576 (1964); **34**, 1544 (1964).
3. Das, A., *Nuovo Cimento*, **18**, 482 (1960).
4. Kadyshevsky, V. G., International Conference on High-Energy Physics at C.E.R.N., Geneva, p. 700, 1962.

Cross-references: ATOMIC CLOCKS, QUANTUM THEORY, RELATIVITY.

TRANSFORMER*

In elementary form, a transformer consists of two coils wound of wire and inductively coupled to each other. When alternating current at a given frequency flows in either coil, an alternating electromotive force (emf) of the same frequency is induced in the other coil. The value of this emf depends on the degree of coupling and the magnetic flux linking the two coils. The coil connected to a source of alternating emf is usually called the primary coil, and the emf across this coil is the primary emf. The emf induced in the secondary coil may be greater than or less than the primary emf, depending on the ratio of primary to secondary turns. A transformer is termed a step-up or a step-down transformer accordingly.

Most transformers have stationary iron alloy cores, around which the primary and secondary coils are placed. Because of the high permeability of iron alloys, most of the flux is confined to the core, and tight coupling between the coils is thereby obtained. So tight is the coupling between the coils in some transformers that the primary and secondary emf's bear almost exactly the same

ratio to each other as the turns in the respective coils or windings. Thus, the turns ratio of a transformer is a common index of its function in raising or lowering potential.

A simple transformer coil and core arrangement is shown in Fig. 1. The primary and secondary coils are wound one over the other on an insulating coil tube or form. The core is laminated to reduce eddy currents. Flux flows in the core along the path indicated, so that all the core flux links both windings. In a circuit diagram, the transformer is represented by the symbol of Fig. 2.

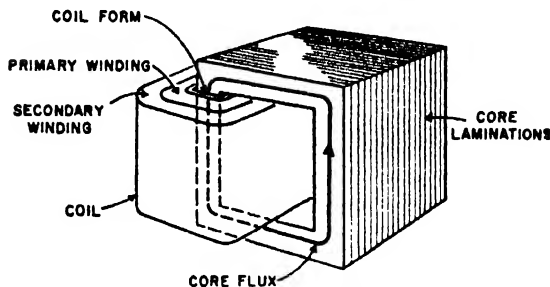


FIG. 1. Transformer coil and core.

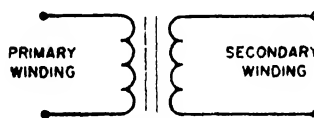


FIG. 2. Simple transformer.

In order for a transformer to deliver secondary emf, the primary emf must vary with respect to time. A dc potential produces no voltage in the secondary winding or power in the load. If both varying and dc potentials are impressed across the primary, only the varying part is delivered to the load. This comes about because the electromotive force e in the secondary is induced in that winding by the core flux ϕ according to the equation

$$e = - \frac{N d\phi}{dt}$$

This equation may be stated in words as follows: The voltage induced in a coil is proportional to the number of turns and to the time rate of change of magnetic flux linking the coil. This rate of change of flux may be large or small. For a given potential, if the rate of change of flux is small, many turns must be used. Conversely, if a small number of turns is used, a large rate of change of flux is necessary to produce a given potential.

Without transformers, modern industry could not have reached its present state of development. The highest potentials which are economically feasible in ac generators are of the order of 20 kV. Transmission of power over long distances is most economical at high potentials which have

* Figures 1 and 2 and information contained in this article are based on the book "Electronic Transformers and Circuits" by Reuben Lee, New York, John Wiley & Sons, Inc., and are used with permission of the publisher.

reached levels of 500 kV and over. The higher the potential of the transmission line, the greater is the amount of power that can be transmitted over a given line conductor. The upper limit of potential is determined by insulation. Insulation research and development have resulted in completely surge-proof transformers, and have made possible power systems which are capable of withstanding lightning surges. At the utilization end of power systems, potential is successively lowered by means of step-down transformers to make power available in safe, useful form. Instrumentation and control of electrical power also require special forms of transformers.

An ideal transformer is defined as one which neither stores nor dissipates energy. Departures from the ideal transformer are caused by:

- (1) Winding resistance and capacitance,
- (2) Leakage inductance (due to flux which does not link both windings),
- (3) Core hysteresis and eddy current losses,
- (4) Magnetizing current.

Factors (1) and (2) above contribute to regulation, or the difference between secondary emf at no-load and full-load. This property is most important with variable load. Although no actual transformer is ideal, some transformers very nearly approach it. For example, in a 50-kVA rectifier transformer winding resistance amounts to 1 per cent, and leakage reactance 3 per cent, of the applied emf; core loss is 0.6 per cent of rated power, and magnetizing current is 2 per cent of rated current. Efficiency (output power divided by input power) is 98.4 per cent for this transformer.

Transformers are needed in electronic apparatus to provide the different values of potential for proper vacuum, gas or solid-state device operation, to insulate circuits from each other, to furnish high impedance to alternating current but low impedance to direct current, to change from one impedance level to another, to connect balanced lines to unbalanced loads, and to maintain or modify wave shape and frequency response at different potentials. Electronic transformers differ from power frequency transformers in the range of impedance levels, frequencies, size and weight. Categories of electronic transformers are: (1) Power, (2) frequency range, and (3) pulse.

Electronic power transformers are generally used to supply rectifiers at potentials ranging from 150 volts to 500 kV. Recent years have seen the widespread use of inverter transformers which convert dc potentials to higher dc potentials in conjunction with semiconductor rectifiers. A typical circuit is that of Fig. 3. Here the transformer output is a rectangular alternating wave which is often rectified again to produce dc output at increased potential.

Frequency range transformers are used in applications where the frequency varies, including audio, video, carrier and control frequencies. Such frequencies vary from a fraction of 1 cps to uhf (300 to 3000 Mc). Transformers may be wide-band or narrow-band in frequency

response, and the core material changes accordingly. In wide-band transformers, the ratio of lowest frequency to highest frequency may be as great as 10^5 . Narrow-band applications use mostly high frequencies and operate over a small percentage of the carrier frequency (e.g., 50 to 55 Mc).

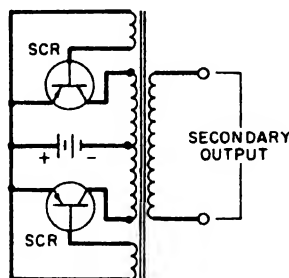


FIG. 3. Solid-state controlled rectifier and transformer from dc to ac inverter.

Pulse transformers are used in radar modulators and computers. In radar applications, the pulses are usually rectangular and occur repetitively. Pulse widths range from 0.1 to 200 μ sec. Peak ratings are large, from 100 to 50 000 kW. Secondary emf may range up to 300 kV for operation of high-power magnetrons and klystrons. Computer transformers are usually small, have ferrite cores and operate from current pulses to drive core matrices.

Advances in transformer technology depend largely on development of new core and insulation materials, conductor arrangements, measuring techniques and methods of application. By these means, transformers come into use at higher frequencies, with better balance, smaller size or higher power than formerly thought possible.

REUBEN LEE

References

- M.I.T. Electrical Engineering Staff, "Magnetic Circuits and Transformers," New York, John Wiley & Sons, 1943.
- Lee, Reuben, "Electronic Transformers and Circuits," Second edition, New York, John Wiley & Sons, 1955.
- "American Standard Requirements, Terminology and Test Code for Distribution, Power and Regulating Transformers and Reactors," C57.12-1956, American Standards Associated, New York.
- "Standard on Low Power Wideband Transformers," No. 111, Institute of Electrical and Electronic Engineers, Box A, Lenox Hill Station, New York, 1964.
- "Proposed Standard on Computer Pulse Transformers" (in preparation), I.E.E.E., New York.
- Bright, R. Louis, Pittman, G. Frank, Jr., and Royer, George H., "Transistors as On-Off Switches in Saturable-Core Circuits," *Elec. Mfg.* (December 1954).

Cross-references: INDUCED ELECTROMOTIVE FORCE, INDUCTANCE, RECTIFIERS.

TRANSISTOR

Few inventions have had as much impact on the electronics industry as the invention of the transistor. Even though the first transistors differed greatly in structure and method of fabrication from those which are in use today, many of the advantages were immediately recognized. Because it was the first workable solid-state amplifier with no physical mechanism which would wear out during its life and because it was relatively simple to fabricate, it immediately became apparent that the transistor would replace a large fraction of the vacuum tubes which were used in radios and electronic systems.

The point-contact transistor¹ was invented by J. Bardeen and W. H. Brattain at Bell Telephone Laboratories in 1948. Its structure (Fig. 1)

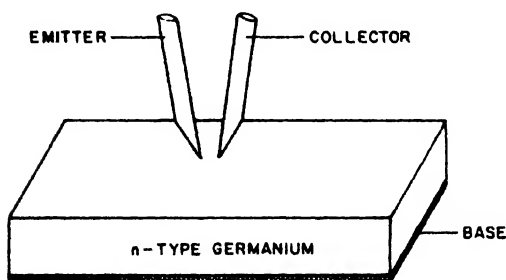


FIG. 1. Point-contact transistor.

consists of a piece of *n*-type germanium about $0.05 \times 0.05 \times 0.005$ inches on which are placed two sharpened points (a beryllium-copper emitter and a phosphor-bronze collector) approximately 0.002 inches apart. Each of these points exhibits rectifying diode type current-voltage characteristics. The operation of the transistor depends on the injection of holes into the *n*-type material through the forward-biased emitter point and the collection at the reverse-biased collector point. The third electrode in the transistor is a low resistance, nonrectifying contact made to the *n*-type germanium and is designated the base electrode. Power gain is achieved in this structure because the collector current increases at a rate equal to two or three times the emitter current and because the reverse-biased collector terminals can be matched in a circuit to a higher impedance load than the input emitter impedance.

The junction transistor^{2,3} exhibits definite advantages over the point-contact transistor and has obsoleted this latter type. In a junction transistor, the rectifying characteristics are obtained within the bulk of the semiconductor crystal by placing different types of impurities at different points in the crystal. The structure of an *n-p-n* junction transistor is illustrated in Fig. 2. The emitter and collector regions have an excess of *n*-type impurities while the base region has an excess of *p*-type impurities. The transition from *p*-type material to *n*-type material is designated as a *p-n* junction. In order to achieve

transistor action it is necessary to have the emitter and collector *p-n* junctions in close physical proximity (0.001 inches or less).

Junction transistors are generally made from germanium or silicon crystals. Germanium technology preceded silicon by approximately five years, but silicon offers several advantages—operation at higher temperatures, lower power consumption, and greater surface stability—so that silicon transistors are rapidly overtaking germanium transistors. Germanium offers only

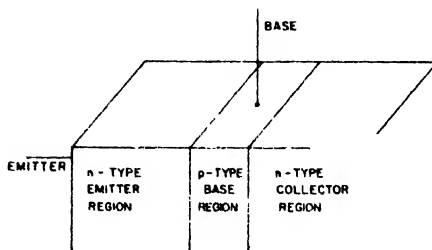


FIG. 2. Structure of *n-p-n* grown-junction transistor.

definite advantages in the high-frequency area because electrons and holes travel faster in germanium than in silicon. Transistors have also been made from other semiconductor materials such as gallium arsenide,⁴ but this technology is relatively new and has not been demonstrated to be as useful. In addition, there are also *p-n-p* transistors where the type of material used for emitter, base, and collector regions is reversed from those shown for the *n-p-n* transistor in Fig. 2. The basic difference is that the polarity of all the operating voltages and currents are opposite for these two types of transistors.

The operation of the *n-p-n* transistor hinges upon the injection of electrons across the forward biased emitter *p-n* junction into the base region of the transistor. Once these electrons are injected across the junction they diffuse across the base and arrive at the reverse-biased collector junction. While the emitter junction acts as a source for these electrons, the reverse-biased collector junction acts as a sink for them. We measure the efficiency of such transistor by the current multiplication factor α_F defined as

$$\alpha_F = \frac{\partial I_C}{\partial I_E} \quad (1)$$

where I_C is the collector current and I_E is the emitter current. In most junction transistors α_F is between 0.95 and 1.00.

The junction transistor can be connected in a circuit so that any one of its three terminals is the common terminal. Figure 3 shows an *n-p-n* transistor connected in the common emitter configuration with typical operating voltages and currents. Note that in this connection a small incremental increase in base current produces a

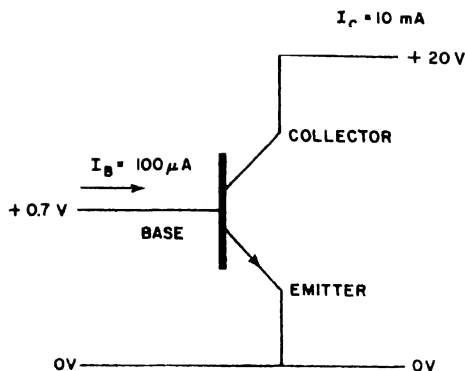


FIG. 3. Circuit symbol and typical potentials for n - p - n silicon junction transistor.

sizeable increase in collector current. This ratio is defined as

$$h_{FE} = \frac{\partial I_C}{\partial I_B} = \frac{\alpha_F}{1 - \alpha_F} \quad (2)$$

Typical values of h_{FE} are between 20 and 200.

There are many types of junction transistors, dependent on the specific method of fabrication chosen. In all cases n -type impurities such as phosphorus, arsenic, or antimony, and p -type impurities such as boron, indium, gallium, or aluminum are introduced into the appropriate regions of the transistor. In a grown-junction transistor,⁵ the impurities are introduced at the time that the silicon or germanium crystal is being fabricated. In an alloy-junction structure,⁶ the impurities are introduced by melting a pellet of the appropriate doping material on the surface of the semiconductor and regrowing the crystal. Diffused junction transistors^{7,8} are produced by placing the semiconductor material in a furnace at an elevated temperature in an atmosphere containing the desired doping element. The doping material then diffuses⁹ into the semiconductor crystal and substitutionally replaces some of the silicon or germanium atoms in the crystal lattice. A region of the opposite type can be created by overcompensating, diffusing into the same region the other type of impurity to a concentration in excess of that previously diffused.

Figure 4 shows the structure for a silicon planar transistor.¹⁰ This transistor is produced by diffusing boron impurities into n -type silicon in order to create the p -type base region and later diffusing phosphorus to create the n -type emitter region. Note that the surface of the transistor is covered by a layer of silicon dioxide (quartz) which acts as a protection for the exposed p - n junctions; such layer assures a very high-reliability transistor.

This transistor is also economical to fabricate because mass processing techniques can be utilized; several hundred identical transistors can be produced on a slice of silicon approximately one inch in diameter. The precise geometrical control achieved by using diffusion

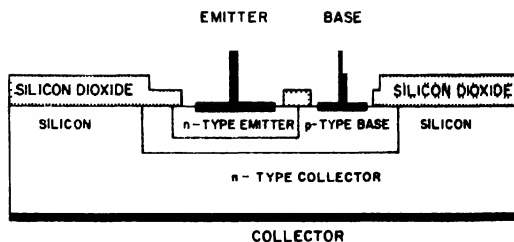


FIG. 4. Cross-sectional view of silicon n - p - n planar transistor.

techniques results in very reproducible electrical characteristics. In order to fabricate a planar transistor we take advantage of the fact that boron and phosphorous do not diffuse through the protecting silicon layer. Hence, it is possible to cover an entire slice of silicon with the oxide and cut holes in the oxide of a geometrical pattern corresponding to that of the desired diffusions. The same techniques can also be used for making integrated circuits.¹¹ In an integrated circuit, the transistors, diodes, resistors, and capacitors are produced on the same slice of silicon by properly controlled diffusion methods.

Transistors are now being used extensively in all electronic equipment. As amplifiers they are useful up to frequencies of several hundred megacycles and at power levels up to 100 watts. However, one of the largest applications of transistors has been as a switch in computers. The electrical characteristics of a transistor are nearly ideal for this type of application in that they closely simulate a relay in its open and closed positions. The fastest and most reliable computers in use today are all transistorized.

L. B. VALDES

References

1. Bardeen, J., and Brattain, W. H., "The Transistor: A Semiconductor Triode," *Phys. Rev.*, **74**, 230-231 (July 15, 1948).
2. Shockley, W., "The Theory of p - n Junctions in Semiconductors and p - n Junction Transistors," *Bell System Tech. J.*, **28**, 435-489 (July, 1949).
3. Shockley, W., Sparks, M., and Teal, G. K., " p - n Junction Transistors," *Phys. Rev.*, **83**, 151-162 (July 1, 1951).
4. Jones, M. E., and Wurst, E. C., Jr., "Recent Advances in Gallium Arsenide Transistors," *IRE Intern. Conv. Record*, **9**, Pt. 3, 26-29 (1961).
5. Wallace, R. L., and Pietenpol, W. J., "Some Circuit Properties and Applications of n - p - n Transistors," *Proc. IRE*, **39**, 753-767 (July 1951).
6. Saby, J. S., "Fused Impurity p - n - p Junction Transistors," *Proc. IRE*, **40**, 1358 (November, 1952).
7. Lee, C. A., "A High-Frequency Diffused Base Germanium Transistor," *Bell System Tech. J.*, **35**, 23-24 (January, 1956).
8. Tanenbaum, M., and Thomas, D. E., "Diffused Emitter and Base Silicon Transistors," *Bell System Tech. J.*, **35**, 1-22 (January, 1956).

9. Fuller, C. S., and Ditzenberger, J. A., "Diffusion of Donor and Acceptor Elements in Silicon," *J. Appl. Phys.*, 27, No. 5, 544-553 (May, 1956).
10. Hoerni, J. A., IRE Electron Devices Meeting, Washington, D.C., October, 1961.
11. Moore, G. E., "Semiconductor Integrated Circuits," "Microelectronics," pp. 262-359, New York, McGraw-Hill Book Company, Inc., 1963.

General References

- Shockley, W., "Electrons and Holes in Semiconductors," Princeton, N.J., D. Van Nostrand Co., 1950.
- Dunlap, W. C., "An Introduction to Semiconductors," New York, John Wiley & Sons, 1957.
- Shive, J. N., Bridgers, H. E., Scaff, J. H., and Biondi, F. J., "Transistor Technology," Vols. I, II, III, Princeton, N.J., D. Van Nostrand Co., 1958.
- Valdes, L. B., "The Physical Theory of Transistors," New York, McGraw-Hill Book Co., 1961.
- Hunter, L. P., "Handbook of Semiconductor Electronics," Second edition, New York, McGraw-Hill Book Co., 1962.
- Spence, E., "Electronic Semiconductors," New York, McGraw-Hill Book Co., 1958.

Cross-references: DIODE (SEMICONDUCTOR), ELECTRON TUBES (RECEIVING TYPE), RECTIFIERS, SEMICONDUCTORS, SOLID-STATE PHYSICS.

TRANSPORT THEORY (Radiative Transfer)*

When light passes through an atmosphere, it may be scattered and absorbed. When neutrons move through a medium, collisions with the nuclei of the material may result in absorption of the neutrons, changes in their direction and energy, and sometimes, by the fission mechanism, production of more neutrons.

These two phenomena are examples of transport processes. The problem of specifying the radiation field in an atmosphere stems back to Rayleigh's investigations on the illumination of a sunlit sky. The astrophysicists refer to this general subject as *radiative transfer* and have studied it for well over half a century. Interest in neutron transport has perforce been of more recent origin.

If light is thought of as consisting of photons, then there is seen to be a very strong similarity between neutron transport and radiative transfer. The former process is both complicated and made more interesting by the possibility of fission in various materials. Other physical phenomena, such as the passage of γ -rays through a medium, possess many characteristics of the two processes that have been mentioned.

Fortunately, it is possible to develop a mathematical structure which encompasses all these

phenomena. The situation is similar to that in classical diffusion theory, where the same mathematical equations may be interpreted, for example, to yield information concerning the distribution of heat in a metal or the flow of one material into another. The equations describing particle transport are, however, of a much more complicated nature than many of those of classical physics and are only now beginning to yield to the techniques of the mathematicians.

All the transport processes described above have the property that the moving particles involved may be thought of as interacting or colliding with fixed centers or nuclei of the material through which they are passing. The moving particles do not collide with each other. The interactions, in the situations which we will consider, are strictly independent and local events—a particle is affected only by a scattering center in its immediate neighborhood. Most important is the fact that a probability for such an interaction may be assigned:

Probability of interaction in moving a distance Δ

$$= \Sigma \Delta + (\text{terms of higher order in } \Delta) \quad (1)$$

The quantity Σ , called the *macroscopic cross section*, is dependent upon the kind and density of the medium, the type of moving particle (photon, neutron, etc.), the particle energy, etc. Determination of Σ is a complicated problem of experimental and theoretical physics.

An interaction or collision may result in a change in the direction of the moving particle (scattering), disappearance of the particle (absorption), or production of new particles of the same kind (fission). Scattering may produce no energy loss (elastic), or it may involve energy loss (inelastic). Such interactions may affect the transport medium, though for most purposes this change may be neglected. Again, the actual physical determination of the result of a collision is frequently a difficult task. Such information, together with the quantity Σ , must be considered for our purposes as already known⁶.

The central problem of transport theory as we view the subject, may now be formulated. Given a physical medium with all parameters, such as the macroscopic cross section and the results of interaction, specified, let a population of particles of given kind, direction, energy, etc., be present in the material at initial time, $t = 0$. Let any internal or external sources of such particles be given. Describe as a function of position, direction, energy, etc., the expected particles population at any time $t > 0$.

Clearly, the transport process is probabilistic in nature. It is customary, because of the difficulty of the subject, to study expected value theory, although some investigations of a more detailed type have been made. Often the term "expected" is dropped in discussion so that "expected flux" becomes simply "flux," etc. It is important to remember that this is done only as a matter of convenience in writing.

* This work was supported by the United States Atomic Energy Commission. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

From Eq. (1) further information concerning the probability of collision may be obtained. Write

$p(x)$ = Probability that a particle moves a distance x without collision (2)

Then,

$p(x + \Delta) = (1 - \Sigma\Delta) p(x) + (\text{terms of higher order in } \Delta)$ (3)

Equation (3) gives

$$\frac{dp}{dx} = -\Sigma p(x)$$

and, if Σ is constant,

$$p(x) = e^{-\Sigma x} \quad (4)$$

This is the well-known exponential law.

To find the average distance a particle moves without collision observe that

$F(x)$ = Probability that the first collision occurs at some $X < x$ (5)

$$1 - e^{-\Sigma x}$$

Then the average or expected value of X is

$$E(X) = \int_0^x x dF(x) = \int_0^x \Sigma x e^{-\Sigma x} dx \quad (6)$$

$$\frac{1}{\Sigma}$$

This average distance between collisions is referred to as the *mean free path*, usually denoted λ .

Suppose the half space $x > 0$ is filled with a medium characterized by constant macroscopic cross section Σ . If a flux of N_0 particles impinges perpendicularly on $x = 0$, then the expected flux at $x = d$ of particles which have made no collision is, from Eq. (4),

$$N(d) = N_0 e^{-\Sigma d} \quad (7)$$

When the collision process involves only absorption, $N(d)$ is the total expected flux at $x = d$, but if scattering and fission processes occur $N(d)$ may be quite different from the total flux. These physical events complicate transport theory greatly.

Some idea of the complexities introduced by the fission process may be obtained by study of a very simple mathematical model in which particles are allowed to move only to the right or left in a rod (i.e., a line segment) of length a . Suppose that in an interaction the colliding particle disappears and two new ones emerge, one moving left and one moving right (binary fission). All particles have the same speed c , and Σ is constant. Finally, assume that the process is such that the average particle density is the same at one time as it is any other, so that time dependence may be

neglected. Denote by $cu(x)$ and $cv(x)$ the right and left fluxes at x . Then?

$$u(x + \Delta) = (1 - \Sigma\Delta)u(x) + \Sigma\Delta u(x) + \Sigma\Delta v(x) + (\text{higher order terms in } \Delta) \quad (8)$$

which leads to

$$\frac{du}{dx} = \Sigma v \quad (9)$$

Similarly

$$-\frac{dv}{dx} = \Sigma u \quad (10)$$

If one left-moving particle per second is introduced at $x = a$, with no source at $x = 0$, then

$$u(0) = 0, \quad cv(a) = 1, \quad (11)$$

and the system of Eqs. (9) through (11) yields

$$cu(x) = \frac{\sin \Sigma x}{\cos \Sigma a}, \quad cv(x) = \frac{\cos \Sigma x}{\cos \Sigma a}, \quad 0 \leq x \leq a \quad (12)$$

These results are obviously quite different from any that would be given by a simple attenuation law such as Eq. (7). Indeed, because of the fission assumed, a collision results in an increase, rather than a decrease, in the particle population. The case $a = \pi/2\Sigma$ is of especial interest. For a rod of that length neither $u(x)$ nor $v(x)$ is defined. Physically, the system is just *critical*. A time-independent population cannot prevail with the source specified. Equations (9) and (10) no longer hold when $a \geq \pi/2\Sigma$.

This observation does not imply that supercritical systems cannot be analyzed. To do so requires explicit introduction of the time variable². Equations (9) and (10) are replaced by

$$\frac{1}{c} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = \Sigma v(x, t) \quad (13)$$

$$\frac{1}{c} \frac{\partial v}{\partial t} - \frac{\partial v}{\partial x} = \Sigma u(x, t) \quad (14)$$

with suitable boundary and initial conditions. For large times it may be shown that the solutions to these equations can be written *approximately* as

$$u(x, t) = u_0(x) e^{\alpha t}, \quad v(x, t) = v_0(x) e^{\alpha t} \quad (15)$$

for some α . When $a > \pi/2\Sigma$, α is positive so that the particle fluxes build up exponentially in time. This exponential increase is observed in actual experiments involving fissionable materials. When the particle population in a system is just sustained without introduction of additional particles (sources) the system is just critical (see CRITICAL MASS).

It is possible to consider other relatively simple mathematical models of transport phenomena and from them to determine much valuable information. Any attempt to solve a "realistic" physical problem, however, usually results in

great difficulties. The general transport equation is of the form

$$\frac{D N}{D t}(\bar{r}, \bar{v}, t) = -\Sigma(\bar{r}, \bar{v}, t) N(\bar{r}, \bar{v}, t) + \int_{\bar{v}'} K(\bar{r}, \bar{v}' \rightarrow \bar{v}, t) \Sigma(\bar{r}, \bar{v}', t) N(\bar{r}, \bar{v}', t) d\bar{v}' + S(\bar{r}, \bar{v}, t) \quad (16)$$

subject to boundary and initial conditions dependent upon the geometry. In Eq. (16), $\frac{D}{D t}$ is the total time derivative, \bar{v} is the velocity, and K is a function that gives the density of particles emerging at velocity \bar{v} from a collision at \bar{r} at time t involving a particle moving at velocity \bar{v}' . Sources are represented by S . It is clear that Eq. (16), while linear, has a much more complicated structure than many of the classical equations of mathematical physics. Little is known of the mathematical properties of its solutions.

The complexity of the general transport operator, coupled in some cases with the urgencies of designing atomic weapons, nuclear reactors, etc., has resulted in a plethora of approximate methods for its treatment.^{3,4,6} It should be pointed out that no such scheme can possibly give satisfactory results without good knowledge, either experimental or theoretical, of the physical parameters involved. Great efforts have been made toward accurate determinations of Σ and K .

A very simple approximation to transport theory may be obtained by considering all particles of the same energy, with K a constant. Truncated expansion in Legendre polynomials of the angular dependence of pertinent functions then yields neutron diffusion theory. The resultant equation is identical in structure to the heat equation, allowing classical solutions to be used in many problems. This approximate theory is valid only under quite stringent conditions, but it often gives surprisingly good results. Variants of simple neutron diffusion theory are numerous, including age theory and n -velocity group theory to account for energy changes, and the P_n -approximation, in which more terms of the Legendre series are retained. Non-constant K may also be included. A method of calculating the integral in Eq. (16) by Gaussian quadrature is closely connected with the P_n -approximation.²

The advent of the high-speed computing machine has made finite difference schemes for solving the transport equation quite feasible. One of the most successful is the so-called S_n -method, characterized by the fact that angular dependence of the flux is assumed to be piecewise linear.¹

Another popular attack on transport problems is via the Monte Carlo method.⁶ There, the particle motion is actually simulated by machine computation. Particles suffer collisions, change direction, lose energy, etc., according to specified probabilities. The machine traces the history of a single particle from its appearance in the system until the particle leaves or is no longer of interest. Events in the particle's history are allowed to

happen randomly, the computing machine deciding on the event according to a preassigned (and presumably physically correct) probability distribution. This method is hence stochastic in nature, although expected values are usually taken after a sufficiently large particle population has been examined. The Monte Carlo technique is often used when the geometry of the problem is complicated.

A completely different approach, of both analytical and computational interest, to the over-all problem of particle transport has been used rather extensively over the last twenty years by the astrophysicists, although it is as yet not well known to workers in neutron theory. This device, called the "method of invariance" or "invariant imbedding," focuses attention on the particles *emergent* from the medium.^{2,7} Equations are derived by considering the relationship between the flux out of a system and the corresponding flux emergent from a slightly larger system. Invariant imbedding has several computational advantages, especially when the emergent flux is the quantity of primary physical interest. The imbedding equations are nonlinear, but in some ways are easier to handle than equations like Eq. (16). At present, the method can be successfully applied only to systems with considerable symmetry.

Transport theory is still a relatively young subject. Much remains to be accomplished, both from the viewpoint of the physicist and from that of the mathematician.

G. MILTON WING

References

1. Carlson, Bengt, "Numerical Solution of Neutron Transport Problems," *Proceedings of the Symposium in Applied Mathematics*, 11, 217-232 (1961).
2. Chandrasekhar, S., "Radiative Transfer," Oxford, Clarendon Press, 1950.
3. Davison, B., "Neutron Transport Theory," Oxford, Clarendon Press, 1957.
4. Kourganoff, V., and Busbridge, I. W., "Basic Methods in Transfer Problems," Oxford, Clarendon Press, 1952.
5. Richtmyer, R. D., "Monte Carlo Methods," *Proceedings of the Symposium in Applied Mathematics*, 11, 190-205 (1961).
6. Weinberg, A. M., and Wigner, E. P., "The Physical Theory of Neutron Chain Reactors," Chicago, The University of Chicago Press, 1958.
7. Wing, G. M., "An Introduction to Transport Theory," New York, John Wiley & Sons, 1962.

Cross-references: CROSS SECTIONS AND STOPPING POWER, COLLISIONS OF PARTICLES, CRITICAL MASS, FISSION, HEAT TRANSFER.

TRANSURANIUM ELEMENTS

Those elements heavier than uranium, element 92. The transuranium elements are all radioactive and have half-lives too short to have existed in nature since their original creation. They were

all discovered and produced by nuclear synthesis. The presently known (1964) transuranium elements have the following names and symbols: 93, neptunium (Np); 94, plutonium (Pu); 95, americium (Am); 96, curium (Cm); 97, berkelium (Bk); 98, californium (Cf); 99, einsteinium (Es); 100, fermium (Fm); 101, mendelevium (Md); 102, (name in question) and 103, lawrencium (Lw).

It is reasonable to assume that elements heavier than lawrencium, 103, will be produced either through the bombardment of the heavier transuranium elements with heavy ions (heavier than helium ions) in an accelerator or in specially tailored thermonuclear devices detonated underground. The half-lives of these elements are expected to be so short as to make conventional chemical identification difficult up to elements 104 and 105 and very difficult beyond these. Perhaps chemical identification can be made in some cases by using simple and fast methods, as for example, those involving migration of gaseous atoms or ions, volatility properties, reactions with surfaces, or gas flow reactions. The historic requirements for the discovery of a new element, namely complete chemical identification and separation from all previously known elements, have already had to be changed with the discovery of the isotope, Lw^{257} , with a half-life of 8 seconds.

The heaviest elements of the periodic system with atomic numbers 89 (actinium) through 103 (lawrencium) are members of the actinide series, analogous to the lanthanide series or rare earths (atomic numbers 57 through 71). An inner electron shell, consisting of fourteen *5f* electrons is filled in progressing across the series.

Since the actinide series was completed with lawrencium, element 103, the succeeding elements yet to be discovered with atomic numbers 104 through 108 should be chemical homologs to the known elements with atomic numbers 72 (hafnium) to 76 (osmium). This analogy would continue in the still heavier transuranium elements, and element 118 would be a rare gas.

Chemically the present transuranium elements, which are all members of the actinide series, are very similar, although the observed differences are those expected and anticipated from their unique position in the periodic system as part of a second rare-earth series. All have trivalent ions, which form inorganic complex ions and organic chelates. Also in common are acid-insoluble trifluorides and oxalates, soluble sulfates, nitrates, chlorides, and perchlorates. Neptunium-plutonium and americium have higher oxidation states in aqueous solution (similar to uranium), but the relative stability of these states to the common trivalent ion becomes progressively less as one proceeds to the higher atomic numbers. This is a direct consequence, indeed an identifying feature, of the actinide role as a second rare-earth type transition series.

One of the most important methods for study and elucidation of chemical behavior of the actinide elements has been ion-exchange chroma-

tography. Adsorption on and elution from ion-exchange columns have made possible the identification and separation of trace quantities of all of the actinides and in particular the transuranium elements. The behavior of each actinide and transuranium element in this respect is very similar to its analogous rare-earth element. This has made it possible to detect as little as one or two atoms when this small a number has been made in some of the transmutation experiments.

The concept of atomic weight in the sense applied to naturally occurring elements is not applicable to the transuranium elements, since the isotopic composition of any given sample depends on its source. In most cases, the use of the mass number of the longest-lived isotope in combination with an evaluation of its availability has been adequate. Good choices at present are neptunium, 237; plutonium, 242; americium, 243; curium, 248; berkelium, 249; californium, 249; einsteinium, 254; and fermium, 255.

Brief descriptions of the transuranium elements follow:

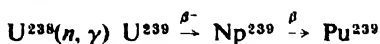
Neptunium. Neptunium (Np, atomic number 93, after the planet Neptune). Neptunium was the first of the synthetic transuranium elements to be discovered; the isotope Np^{239} was produced by McMillan and Abelson in 1940 at Berkeley, California, as the result of the bombardment of uranium with cyclotron-produced neutrons. The isotope Np^{237} (half-life of 2.2×10^6 years) is currently obtained in kilogram quantities as a by-product from nuclear reactors in the production of plutonium. Trace quantities of the element are actually found in nature owing to transmutation reactions in uranium ores produced by the neutrons which are present.

Neptunium metal has a silvery appearance, is chemically reactive, melts at 640°C and exists in at least three structural modifications: α -neptunium, orthorhombic, density = 20.45 g/cm^3 ; β -neptunium (above 278°C), tetragonal, density (313°C) = 19.36 g/cm^3 ; γ -neptunium (above 500°C), cubic, density (600°C) = 18.0 g/cm^3 .

Neptunium gives rise to four ionic oxidation states in solution: Np^{+3} (pale purple), analogous to the rare earth ion Pm^{+3} , Np^{+4} (yellow green), NpO_2^+ (green blue), and NpO_2^{++} (pale pink). These latter oxygenated species are in contrast to the rare earths which exhibit only simple ions of the (II), (III), and (IV) oxidation states in aqueous solution. The element forms tri- and tetrahalides such as NpF_3 , NpF_4 , NpCl_4 , NpBr_3 , NpI_3 , and oxides of various compositions such as are found in the uranium-oxygen system, including Np_2O_7 and NpO_2 .

Plutonium. Plutonium (Pu, atomic number 94, after the planet Pluto). Plutonium was the second transuranium element to be discovered; the isotope Pu^{239} was produced in 1940 by Seaborg, McMillan, Kennedy, and Wahl at Berkeley, California, by deuteron bombardment of uranium in the 60-inch cyclotron. By far of greatest importance is the isotope Pu^{239} (half-life of 24,400 years) which is fissionable with thermal neutrons and produced in extensive quantities in

nuclear reactors from the abundant nonfissionable uranium isotope U^{238} :



Plutonium has assumed the position of dominant importance among the transuranium elements because of its successful use as an explosive ingredient in nuclear weapons and the place which it holds as a key material in the development of industrial utilization of nuclear energy, one pound being equivalent to about 10 000 kWh of heat energy. In certain nuclear reactors called breeder reactors, it is possible to create more new plutonium from U^{238} than plutonium consumed in sustaining the fission chain reaction. Because of this, plutonium is the key to unlocking the enormous energy reserves in the nonfissionable isotope U^{238} .

Plutonium also exists in trace quantities in naturally occurring uranium ores. It is formed in much the same manner as neptunium, by irradiation of natural uranium with the neutrons which are present.

Plutonium metal can be prepared, in common with neptunium and uranium, by reduction of the trifluoride with alkaline-earth metals. The metal has a silvery appearance, is chemically reactive, and melts at 640°C. It exhibits six crystalline modifications: α -plutonium, orthorhombic, below 122°C; β -plutonium, monoclinic, 122 to 206°C; γ -plutonium, 206 to 319°C; δ -plutonium, face-centered cubic, 319 to 451°C; δ' -plutonium, tetragonal, 451 to 476°C and ϵ -plutonium, body-centered cubic, 476°C up to the melting point.

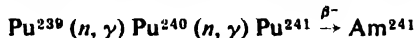
Plutonium also exhibits four ionic valence states in aqueous solutions: Pu^{+3} (blue lavender), Pu^{+4} (color unknown), AmO_2^{+2} (light tan), and a fluoride complex of the +4 state. The trivalent state is highly stable and difficult to oxidize. AmO_2^{+2} , like plutonium, is unstable with respect to disproportionation into Am^{+3} and AmO_2^{+2} . The ion Am^{+4} may be stabilized in solution only in the presence of very high concentrations of fluoride ion, and tetravalent solid compounds are well known. There is some evidence that Am^{+2} has been prepared in tracer experiments at very low concentrations; this would be very similar to the analogous lanthanide, europium, which can be reduced to the divalent state.

Americium forms binary compounds with oxygen: PuO , PuO_2 , and intermediate oxides of variable composition; with the halides: PuF_3 , PuF_4 , $PuCl_3$, $PuBr_3$, PuI_3 ; with carbon, nitrogen and silicon; PuC , PuN , $PuSi_2$; in addition oxyhalides are well known: $PuOCl$, $PuOBr$, $PuOI$.

Because of the high rate of emission of alpha particles, and the physiological fact that the element is specifically absorbed by bone marrow, plutonium, as well as all of the other transuranium elements except neptunium, are radiological poisons and must be handled with special equipment and precautions.

Americium. Americium (Am, atomic number 95, after the Americas). Americium was the fourth transuranium element to be discovered; the isotope Am^{241} was identified by Seaborg, James, Morgan and Ghiorso late in 1944 at the wartime Metallurgical Laboratory (now the Argonne National Laboratory) of the University of Chicago as the result of successive neutron capture reac-

tions by plutonium isotopes in a nuclear reactor:



Americium is produced in kilogram quantities. Since the isotope Am^{241} can be prepared in relatively pure form by extraction as a decay product over a period of years from plutonium containing Pu^{241} , this isotope is used for much of the chemical investigation of this element. Better suited is the isotope Am^{243} owing to its longer half-life (8.8×10^3 years as compared to 470 years for Am^{241}). A mixture of the isotopes Am^{241} , Am^{242} , and Am^{243} can be prepared by intense neutron irradiation of Am^{241} according to the reactions $Am^{241}(n, \gamma) Am^{242}(n, \gamma) Am^{243}$. Nearly isotopically pure Am^{243} can be prepared by a sequence of neutron bombardments and chemical separations as follows: Neutron bombardment of Am^{241} yields Pu^{242} by the reactions $Am^{241}(n, \gamma) Am^{242} \xrightarrow{\beta^-} Pu^{242}$; after chemical separation the Pu^{242} can be transformed to Am^{243} via the reactions $Pu^{242}(n, \gamma) Pu^{243} \xrightarrow{\beta^-} Am^{243}$, and the Am^{243} can be chemically separated. Fairly pure Pu^{242} can be prepared more simply by very intense neutron irradiation of Pu^{239} as the result of successive neutron-capture reactions.

Americium can be obtained as a silvery white reactive metal by reduction of americium trifluoride with barium vapor at 1000 to 1200°C. It appears to be more malleable than uranium or neptunium and tarnishes slowly in dry air at room temperature. The density is 13.67 g/cm³ with a melting point at 994°C.

The element exists in four oxidation states in aqueous solution: Am^{+3} (light salmon), AmO_2^{+2} (color unknown), AmO_2^{+2} (light tan), and a fluoride complex of the +4 state. The trivalent state is highly stable and difficult to oxidize. AmO_2^{+2} , like plutonium, is unstable with respect to disproportionation into Am^{+3} and AmO_2^{+2} . The ion Am^{+4} may be stabilized in solution only in the presence of very high concentrations of fluoride ion, and tetravalent solid compounds are well known. There is some evidence that Am^{+2} has been prepared in tracer experiments at very low concentrations; this would be very similar to the analogous lanthanide, europium, which can be reduced to the divalent state.

Americium dioxide, AmO_2 , is the important oxide; Am_2O_3 and, as with previous actinide elements, oxides of variable composition between $AmO_{1.5}$ and AmO_2 are known. The halides AmF_3 , AmF_4 , $AmCl_3$, $AmBr_3$, and AmI_3 have also been prepared.

Curium. Curium (Cm, atomic number 96, after Pierre and Marie Curie). Although curium comes after americium in the periodic system, it was actually known before americium and was the third transuranium element to be discovered. It was identified by Seaborg, James and Ghiorso in the summer of 1944 at the wartime Metallurgical Laboratory in Chicago as a result of helium-ion bombardment of Pu^{239} in the Berkeley, California, 60-inch cyclotron. It is of special interest because it is in this element that the first

half of the transition series of actinide elements is completed.

The isotope Cm^{242} (half-life 162.5 days) produced from Am^{241} by the reactions $\text{Am}^{241}(n, \gamma) \text{Am}^{242} \xrightarrow{\beta^-} \text{Cm}^{242}$ has been used for much work with macroscopic quantities, although this is difficult due to the extremely high specific alpha activity. An excellent isotope for the investigation of curium in weighable amounts is Cm^{244} because it has a half-life of 19 years and can be prepared in fairly pure form by a sequence of neutron bombardments and chemical separations as follows. The neutron bombardment of Am^{241} yields Pu^{242} by the reactions $\text{Am}^{241}(n, \gamma) \text{Am}^{242} \xrightarrow{\beta^-} \text{Pu}^{242}$; after chemical separation the Pu^{242} can be transformed to Am^{243} via the reactions $\text{Pu}^{242}(n, \gamma) \text{Pu}^{243} \rightarrow \text{Am}^{243}$. Fairly pure Pu^{242} can be prepared more simply by very intense neutron irradiation of Pu^{239} as the result of successive neutron-capture reactions. The Am^{243} can be chemically separated and finally transmuted to Cm^{244} through neutron bombardment by the reactions $\text{Am}^{243}(n, \gamma) \text{Am}^{244} \rightarrow \text{Cm}^{244}$. Further neutron bombardment of Cm^{244} produces higher-mass isotopes of longer half-life which are even better for use in the investigation of curium.

Curium metal resembles americium metal in crystal structure but melts at the considerably higher temperature of 1340°C. It can be prepared by heating curium trifluoride with barium vapor at 1350°C.

The main oxidation state in aqueous solutions which has been definitely identified is Cm^{+3} , although a solid, black CmO_2 has been prepared. Cm_2O_3 and CmF_3 , both white compounds, have also been characterized. The general trend of decreasing stability of higher oxidation states with increasing atomic number would indicate that simple ions of curium higher than curium (III) will probably not be stable in aqueous solutions. It has further been demonstrated that curium (III) cannot be reduced in aqueous solutions.

The experimentally determined magnetic susceptibility of CmF_3 is in good agreement with the value expected for the electronic structure $5f^7$ in curium (III), assuming a Russell-Saunders coupling scheme.

Berkelium. Berkelium (Bk, atomic number 97, after Berkeley, California). Berkelium, the eighth member of the actinide transition series, was discovered in December, 1949, by Thompson, Ghiorso, and Seaborg and was the fifth transuranium element synthesized. It was produced by cyclotron bombardment of Am^{241} with helium ions at Berkeley, California.

The chemical properties of berkelium have been studied partly through the use of tracer amounts. The pale orange dioxide has been prepared by igniting the sulfate at 1200°C. These studies have demonstrated that berkelium exists in aqueous solutions in two oxidation states, berkelium (III) and berkelium (IV). The chemistry and solubility of compounds appears to closely follow the other transuranium elements.

The existence of Bk^{249} with a half-life of about

300 days makes it feasible to isolate berkelium in weighable amounts so that its properties can be investigated with macroscopic quantities. This isotope can be prepared by the intense neutron bombardment of Cm^{244} as the result of the capture of successive neutrons by the reactions $\text{Cm}^{244}(n, \gamma) \text{Cm}^{245}(n, \gamma) \text{Cm}^{246}(n, \gamma) \text{Cm}^{247}$

$(n, \gamma) \text{Cm}^{248}(n, \gamma) \text{Cm}^{249} \xrightarrow{\beta^-} \text{Bk}^{249}$.

Californium. Californium (Cf, atomic number 98, after the state and University of California). Californium, the sixth transuranium element to be discovered, was produced by Thompson, Street, Ghiorso, and Seaborg in January, 1950, by helium-ion bombardment of microgram quantities of Cm^{242} in the Berkeley 60-inch cyclotron. Californium (III) is the only ion stable in aqueous solutions, all attempts to reduce or oxidize californium (III) having failed. The existence of the isotopes of Cf^{249} , Cf^{250} , Cf^{251} , and Cf^{252} makes it feasible to isolate californium in weighable amounts so that its properties can be investigated with macroscopic quantities. Both the solid trichloride and oxychloride have been prepared. The isotope Cf^{249} results from the beta decay of Bk^{249} while the heavier isotopes are produced by intense neutron irradiation by the reactions $\text{Bk}^{249}(n, \gamma) \text{Bk}^{250} \xrightarrow{\beta^-} \text{Cf}^{250}$ and $\text{Cf}^{249}(n, \gamma) \text{Cf}^{250}$ followed by $\text{Cf}^{250}(n, \gamma) \text{Cf}^{251}(n, \gamma) \text{Cf}^{252}$.

Einsteinium. Einsteinium (Es, atomic number 99, after Albert Einstein). Einsteinium, the seventh transuranium element to be discovered, was identified by Ghiorso *et al.* in December, 1952, in the debris from a thermonuclear explosion in work involving the University of California Radiation Laboratory, the Argonne National Laboratory, and the Los Alamos Scientific Laboratory. The isotope produced was the 20-day Es^{253} , originating from beta decay of U^{253} and daughters. Its chemical properties have been studied solely with tracer amounts and indicate that the (III) oxidation state may be the only one which exists in aqueous solution. Existence of a relatively long-lived isomeric form of Es^{254} makes it possible to isolate this element in weighable amounts. The isotope Es^{253} and heavier isotopes can be produced by intense neutron irradiation of lower elements such as plutonium, by a process of successive neutron capture interspersed with beta decays until these mass numbers and atomic numbers are reached.

Fermium. Fermium (Fm, atomic number 100, after Enrico Fermi). Fermium, the eighth transuranium element to be discovered, was identified by Ghiorso *et al.* early in 1953 in the debris from a thermonuclear explosion in work involving the University of California Radiation Laboratory, the Argonne National Laboratory, and the Los Alamos Scientific Laboratory. The isotope produced was the 16-hour Fm^{255} , originating from the beta decay of U^{255} and daughters. Since this is the longest-lived presently known fermium isotope, it seems unlikely that this element can be isolated in weighable amount. Its chemical properties have been studied solely

with tracer amounts, and in normal aqueous media only the (III) oxidation state appears to exist. The isotope Fm^{254} and heavier isotopes can be produced by intense neutron irradiation of lower elements, such as plutonium, by a process of successive neutron capture interspersed with beta decays until these mass numbers and atomic numbers are reached.

Mendelevium. Mendelevium (Md, atomic number 101, after Dmitri Mendeleev). Mendelevium, the ninth transuranium element to be discovered, was first identified by Ghiorso, Harvey, Choppin, Thompson, and Seaborg in early 1955 as a result of the bombardment of the isotope Es^{253} with helium ions in the Berkeley 60-inch cyclotron. The isotope produced was presumably Md^{256} which apparently decays by electron capture to Fm^{256} , which in turn decays predominantly by spontaneous fission with a half-life of about 3 hours. This first identification was notable in that only of the order of one to three atoms per experiment were produced. The extreme sensitivity for detection depended on the fact that its chemical properties could be accurately predicted as *eka-thulium* and there was a high sensitivity for detection because of the spontaneous fission decay. The chemical properties have been investigated solely by the tracer technique and seem to indicate that the predominant oxidation state in aqueous solution is the (III) state. There seem to be no isotopes of sufficiently long half-life to make it possible to isolate this element in weighable quantity.

Element 102. An isotope of element 102, 102^{254} with a half-life of about 3 seconds, was discovered by Ghiorso, T. Sikkeland, J. R. Walton, and Seaborg at the University of California in Berkeley in 1958. The element was produced by the bombardment of Cm^{246} with C^{12} ions accelerated in the heavy ion linear accelerator. The isotope was identified through chemical identification of the known daughter Fm^{250} , which was separated by a recoil technique from its alpha-decaying parent.

Earlier in 1957, a team of scientists from the United States, England and Sweden announced the apparent discovery of an isotope of element 102 as a result of research performed at the Nobel Institute for Physics in Stockholm. The name nobelium was suggested by this group for element 102. However, neither experiments at the University of California in Berkeley, which were more sensitive than the Stockholm work, nor related experiments performed in the USSR, have confirmed this apparently erroneous discovery. Although the name nobelium for element 102 will undoubtedly have to be rejected, no suggestion for a new name has yet been made.

Lawrencium. Lawrencium (Lw, atomic number 103, after Ernest O. Lawrence). Lawrencium was discovered in 1961 by A. Ghiorso, T. Sikkeland, A. E. Larsh, and R. M. Latimer using the heavy-ion linear accelerator at the University of California at Berkeley. A few micrograms of a mixture of Cf^{249} , Cf^{250} , Cf^{251} , and Cf^{252} were bombarded with B^{10} and B^{11} ions to produce Lw^{257} which has a

half-life of 8 seconds. It was determined that this isotope decays by the emission of an 8.6-MeV alpha particle.

At present, because of the short half-life of lawrencium, it has not been possible to perform a chemical identification and the discovery rests on nuclear evidence. Lawrencium is the fourteenth and last member of the actinide series and therefore should be chemically similar to the lighter transuranium elements.

GLENN T. SEABORG

Cross-references: ATOMIC PHYSICS; CYCLOTRON; ELEMENTS, CHEMICAL; ISOTOPES; NUCLEAR REACTIONS; NUCLEAR REACTORS; RADIOACTIVITY; RARE EARTHS.

TUNNELING

Tunneling is a quantum mechanical process without a classical analog. An electron (or other quantum mechanical particle) incident upon a potential barrier whose height is larger than the kinetic energy of the electron will penetrate (tunnel) a certain distance into the barrier. This is most easily visualized by considering the one-dimensional Schrödinger equation for the wave function $\psi(x)$ of such an electron:

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V(x)\psi = E\psi \quad (1)$$

If V varies relatively slowly with distance, the solution of Eq. (1) can be approximated by

$$\psi \approx \psi_0 \exp \left[\pm i \frac{\sqrt{2m}}{\hbar} \int \sqrt{V(x) - E} dx \right] \quad (2)$$

The non classical case is the one where the electron energy E is smaller than the potential $V(x)$. Then ψ is no longer an oscillating wave, but it is real and decays exponentially with distance into the barrier. The penetration probability through a given barrier is equal to the ratio of the probability density $|\psi(x)|^2$ at the exit from the barrier to its value at the entrance. Equation (2) shows that this probability is exponentially dependent on barrier height and thickness. The tunneling takes place at constant electron energy as there is no scattering involved.

A more accurate solution of Eq. (1), which adequately described many real situations, may be calculated by the Wentzel, Kramers, Brillouin (WKB) approximation. The tunneling probability through a barrier then becomes

$$P \approx \exp \left[-2 \int_{x_1}^{x_2} \frac{\sqrt{2m}}{\hbar} \sqrt{V(x) - E} dx \right] \quad (3)$$

An idea of the magnitude of the tunneling probability may be obtained by calculating the value of Eq. (3) for a rectangular barrier 1 eV higher than the particle energy and 25 Å wide. The result for an electron is $P \approx 10^{-11}$; for a proton, $P \approx 10^{-400}$.

Tunneling can occur in many physical situations. The simplest case is that of FIELD EMISSION.

This is the emission of electrons from a solid by the application of very high fields. The energy-band diagram for such a system is shown in Fig. 1.

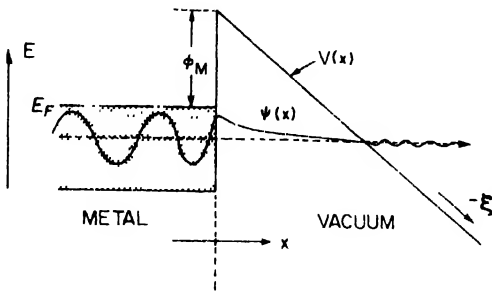


FIG. 1. Simplified energy-band diagram for field emission from a metal. $\psi(x)$ represents the wave function of an electron with energy E .

The electrons in a metal surrounded by vacuum can be regarded as an ensemble of quasi-free electrons held in a potential well by the positive charges of the metal ions. The depth of the well is called the electron affinity χ . The potential well is filled with electrons up to the Fermi level (E_F) and consequently electrons at this level "see" a barrier $\phi_M = \chi - E_F$ (work function) surrounding them. They can tunnel a short distance into the walls of the potential well, but of course cannot escape from it.

If a high positive field ξ is applied to the metal (e.g., by applying a potential difference between the metal under consideration and an adjacent piece of metal), the potential energy distribution is that shown in the Figure. There now exists a barrier through which, under appropriate conditions, the electrons may tunnel. The minimum height of the barrier is ϕ_M , and its width is approximately $\frac{\phi_M}{\xi}$ so that the current becomes

$$I \propto P \approx \exp \left[-2 \frac{2}{3} \frac{\sqrt{2m}}{\hbar} \sqrt{\phi_M} \frac{\phi_M}{\xi} \right] \approx \exp \left[-10^8 \frac{\phi_M^2}{\xi} \right] \quad (4)$$

if ϕ_M is given in units of electron volts and ξ in units of volts per centimeter. Since ϕ_M generally has a value between 2 and 5 eV, an electric field of the order of 10^7 volts/cm is required before appreciable numbers of electrons will be emitted into vacuum.

Figure 1 is a somewhat simplified picture of an actual metal-vacuum interface. An electron outside the metal experiences a polarization force towards the metal, the so-called image force. While this is not a true potential, it may nevertheless be included in the potential energy diagram as a rounding-off of the well edge. When the field is applied, this will mean that the barrier is lowered and so the tunneling probability is increased.

Particularly at high fields this will cause a departure from the simple current relationship described above. If these effects are included in the calculation, good agreement can be obtained with experimental results.

Current flow at a metal-insulator interface may be treated in the same way as that at a metal-vacuum boundary. The potential barrier height is now given by the distance from the metal Fermi level to the bottom of the insulator conduction band, which is usually smaller than the work function. Also the dielectric constant of the insulator will have to be taken into account (e.g., in the image force). These modifications do not change the basic current-voltage relationship [Eq. (4)] which has also been verified experimentally.

In case of a very thin insulator layer ($\sim 60\text{\AA}$) bounded on both sides by metallic regions, high fields are no longer required for tunneling. The band structure may be approximated by the square potential barrier discussed initially. The top of the barrier is again formed by the conduction band of the insulator, modified appropriately by image force considerations. In equilibrium, there will be equal numbers of electrons tunneling in both directions through the insulator; no net current flows.

When a potential difference is applied, one tunneling direction is favored and net current flows. For small potential differences (much less than the barrier heights) the tunneling probability does not vary with applied field and the current flowing will be proportional to the difference in number of electrons available on the two sides of the barrier at the same energies. These numbers are in turn closely related to the density of electronic states in the metals at these energies. The tunneling current may therefore be used to investigate the density of states in certain materials, where this quantity changes rapidly with energy.

This technique for determining the density of states has been put to particularly good use in the case of superconductors. The fact that the superconductor has an energy gap in the density-of-states function at the Fermi level leads to nonlinearities and negative resistance regions in the current-voltage characteristics. This has become the most accurate method of measuring the energy gap of superconductors as a function of temperature, magnetic field and other variables. More complicated phenomena, such as simultaneous tunneling of two electrons as a pair, have also been observed.

Tunneling out of a three-dimensional well is treated similarly to tunneling through the one-dimensional barrier. The exact exponential dependence will be different from Eq. (2) and depend on the form of the barrier. One type of three-dimensional well is formed by an impurity in a semiconductor or insulator which may form a bound state in the region of the forbidden gap. An electron can only tunnel out of such a state when a high field is applied, analogous to field emission from metals.

A different type of three-dimensional potential

well is found in the atomic nucleus. A combination of short-range nuclear attraction and Coulomb repulsion forms a potential barrier of the type shown in Fig. 2. Many heavy radioactive nuclei contain α -particles with high enough kinetic energy to tunnel through the barrier (α -decay). Because of the much heavier mass and high barrier energy, the barrier must be considerably thinner than for an electron ($\sim 10^{-12}$ cm). Again calculations for a Coulomb barrier agree well with experimental observations of the energy dependence of the decay time (Geiger-Nuttall relation).

Electron tunneling is also observed in semiconductors (Zener tunneling), but in this case the situation cannot be represented by the kind of barrier discussed previously. The band structure of a semiconductor under a large applied field is shown in Fig. 3. There exists no set of real energy states connecting the two regions, and an electron must always make a discontinuous step in passing from the valence band to the conduction band. Still the tunneling probability may be calculated with an approximation equivalent to that used for the simple barrier, and except for a numerical factor, the result is the same as Eq. (4) with ϕ_M replaced by E_g .

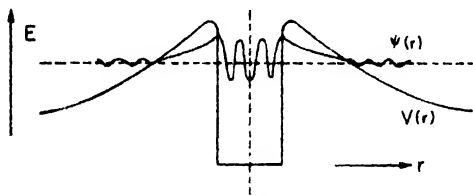


FIG. 2. Model of the nucleus, with $\psi(r)$ the wave function of an α -particle.

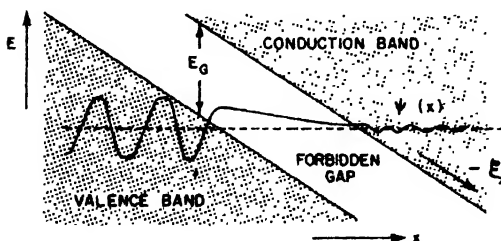


FIG. 3. Band structure of a semiconductor or insulator under an applied field.

Conduction of this type is observed in very narrow p - n junctions under reverse bias (Zener breakdown) or in insulators where there are no free carriers available. In the former case, there is a large built-in field even at zero external voltage which is produced by the difference in the electrochemical potentials of the n - and p -regions. A small additional voltage (1 to 2 volts) raises the field to the value required for tunneling.

In still narrower p - n junctions ("tunnel diodes") tunneling readily takes place even where no applied voltage exists. As in the case of the very thin insulator, the electrons tunneling in opposite

directions then just cancel one another out. The current-voltage curve of such a junction is drawn in Fig. 4. The band structure configurations at two values of forward bias are also shown. Net

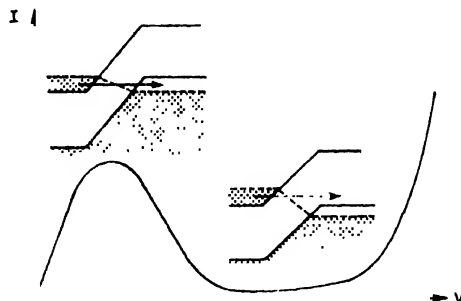


FIG. 4. Current-voltage characteristic of a tunnel diode. The two inserts depict the band structure at the maximum tunnel current and at the valley where tunneling is no longer possible.

tunneling current will flow in either direction of applied voltage, as long as there are electrons on one side opposite empty states of the same energy on the other side. Under large forward current this condition is no longer fulfilled, as the bottom of the conduction band on the n -type side comes opposite the forbidden energy gap on the p -type side. This means that the current passes through a maximum value and then decreases towards zero with increasing forward bias. In actual diodes, the current never decreases completely to zero since the electrons can interact with impurity states in the forbidden gap, dissipating their excess energy so they can drop into the valence band states on the p -type side. At still larger bias, the conventional forward diode current becomes dominating.

DIETRICH MEYERHOFFER

References

- Harrison, W. A., *Phys. Rev.*, **123**, 85 (1961).
- Franz, W., in Flüge, S., Ed., "Handbuch der Physik," Vol. XVII, pp. 201-219, Berlin, Springer-Verlag 1956.
- Fisher, J. C., and Giaever, I., *J. Appl. Phys.*, **32**, 172 (1961).
- International Conference on the Science of Superconductivity, *Rev. Mod. Phys.*, **36**, 200-225 (1964).
- Breit, G., in Flüge, S., Ed., "Handbuch der Physik," Vol. XLI/1, pp. 355ff, Berlin, Springer-Verlag, 1959.
- Chynoweth, A. G., Feldman, W. L., Lee, C. A., Logan, R. A., Pearson, G. L., and Aigrain, P., *Phys. Rev.*, **118**, 425 (1960).
- Kane, E. O., *J. Appl. Phys.*, **32**, 83 (1961).

Cross-references: ENERGY LEVELS, FIELD EMISSION, RADIOACTIVITY, SCHRÖDINGER EQUATION, SEMICONDUCTORS, SOLID-STATE PHYSICS, SOLID-STATE THEORY, SUPERCONDUCTIVITY.

U

ULTRASONICS

Ultrasonic Waves. Ultrasonic waves are sound waves above the frequency normally detectable by the human ear, that is above about 20 kc/sec. The particles of matter transmitting a *longitudinal* wave move back and forward about mean positions in a direction parallel to the path of the wave. Alternate compressions and rarefactions in the transmitting material exist along the wave propagation direction. In *shear* waves, the particles move perpendicularly to the direction of wave propagation. In surface waves in seismological studies and in waves through thin stock, the *Rayleigh* and *Lamb* waves respectively, the particles undergo much more complicated vibratory motions than in longitudinal and transverse waves (see WAVE MOTION).

TABLE 1. APPROXIMATE SOUND VELOCITIES

	In Water (m/sec)	In Steel (m/sec)
Longitudinal	1.5×10^3	6×10^3
Transverse	Cannot be supported	3×10^3
Rayleigh	..	3×10^3

In most practical applications of ultrasonics, pulses or packets containing a number of oscillation cycles are sent through the solid or liquid under investigation. A longitudinal wave pulse, when incident on the boundary between two materials having different sound velocities, is transformed into reflected and refracted shear and longitudinal waves. Snell's law governs the angles of REFLECTION and REFRACTION for both types of waves. It states that

$$\frac{\sin \theta}{V} = \text{Constant}$$

where θ is the angle the beam makes with a normal to the intervening surface and V is the sound velocity. Therefore, in Fig. 1,

$$\text{Constant} = \frac{\sin \theta_1}{V_{L1}} = \frac{\sin \theta_2}{V_{L2}} = \frac{\sin \phi_1}{V_{S1}} = \frac{\sin \phi_2}{V_{S2}}$$

Transducers. In general, a transducer is a device to convert electrical energy into mechanical vibrations which are then directed into the specimen. The transducer also commonly reconverts received mechanical oscillations back again into

electrical signals for amplification and recognition. Transducers are usually piezoelectric, ferroelectric or magnetostrictive in nature.

The application of a voltage across a piezoelectric crystal causes it to deform with an amplitude of deformation proportional to the voltage. Reversal of the voltage causes reversal of the mechanical strain. Until recently, the only piezoelectric transducer in general use was quartz, but now several synthetic ceramic materials are employed.

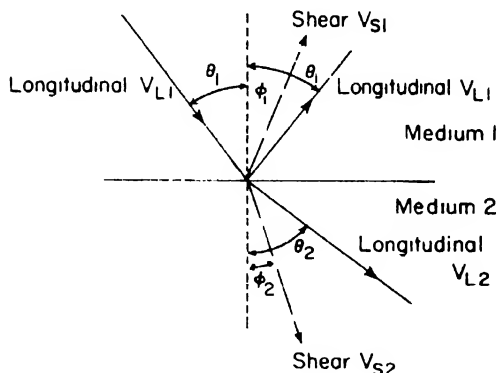


FIG. 1.

Ferroelectric crystals are also electrostrictive. Barium titanate has an electrical mechanical conversion efficiency about one hundred times that of quartz. Unlike the piezoelectric mode of oscillation, in ferroelectric crystals application of a voltage in either direction across the crystal causes expansion of the crystal. This mode can however be converted to the piezoelectric by biasing the expansion in one direction either by application of a strong dc field or more commonly by cooling the ferroelectric crystal through its Curie temperature while it is under the influence of a strong electric field (order of 10^6 volts/m).

Transducer crystals are normally cut to a resonant frequency, the thickness being one-half the acoustic wavelength. A bond between the crystal transducer and the specimen matches the acoustic impedance and carries the acoustic power into the latter. Backing layers may be fixed to the rear surface of the transducer. These layers are chosen to reflect power forward into

the crystal and specimen in some technological applications. They may also be chosen to absorb power so as not to complicate signals received in material testing applications.

Ultrasonic Testing. Most ultrasonic test equipment employs pulses of high-frequency sound (>1 Mc/sec), the pulse width being adjustable between 0.1 and 1.0 μ sec and the repetition rate between 60 and 1000 pulses per second. Figure 2 shows a typical block diagram.

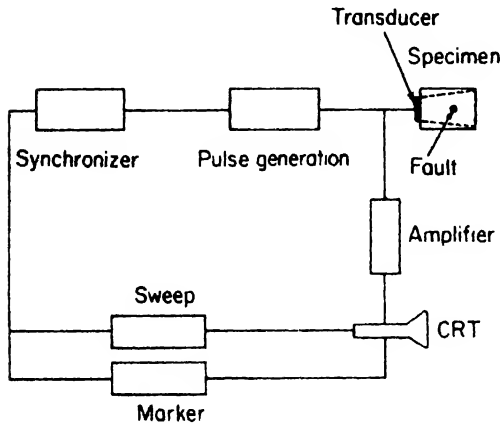


FIG. 2.

The synchronizer triggers a pulse in the generation circuit. This is converted to an acoustic pulse by the transducer crystal. The synchronizer also starts the sweep circuit of the cathode ray tube (CRT) and the marker circuit, the latter making marker pips along the time base. Separate transducer crystals may be used for transmission and reflection, or one crystal may be made to carry out both functions. The echo from the back face of the specimen block arrives twice the transmission time after the initial pulse. The marker pips can then be calibrated in depth. A fault will show up as an extra echo whose depth into the specimen can be read from the marker pip which coincides with its position on the screen of the CRT.

The transducer is coupled to the specimen under test by an oil or adhesive. In high-speed testing or in cases where rough surfaces are encountered, the specimen may be immersed in a liquid or a stream of liquid passed between the transducer and the specimen in order to give sonic coupling.

Examples of Ultrasonic Testing. Solid objects of thickness greater than about half an inch may be tested for inhomogeneities. Special techniques also exist for sheet testing. Heavy forgings, crank shafts, rails, concrete structures and a host of other manufactured objects are now regularly scanned for porosity or internal faults. The frequencies employed are normally from 1 to 5 Mc/sec for steel objects, about 0.5 Mc/sec for many plastics and 0.1 Mc/sec for concrete.

In a medical application of ultrasonic techniques, the subject is immersed in water and the

transducer is scanned in a circular path round the axis of the torso or limb under detailed examination. Echoes are recorded on the CRT which is also circularly scanned (like the P.P.I. radar system). A long persistence screen gives a complete record of the scanned portion of the body. Gallstones, tumors, cancers and other pathological conditions are revealed.

Sophisticated scanning arrays are used in SONAR underwater detection. Frequencies must be low because of the high attenuation in water (at 100 kc/sec attenuation ≈ 60 decibels/km), and this limits definition. Sea noises, fish and underwater refraction by layers at different temperatures introduce difficulty of interpretation into the echo patterns. Depth recording is now accomplished by the use of many sending and receiving transducers coupled to a computing system which interprets differences of arrival times in terms of variation of path length and ultimately of depth. Continuous records of the ocean bed or the depth of surface ice can be made.

Systems similar in principle to Sonar are also employed to give continuous records of manufactured plate thickness and to control the filling of enclosed tanks.

Examples of the Use of Ultrasonic Power. Longitudinal ultrasonic waves can be produced at sufficient intensity to cause changes in materials of industrial importance. In many cases, standing waves are set up in a containing vessel with resulting periodic variation of vibration amplitude. In almost all cases ceramic transducers are used. Emulsifying, mixing, dispersing, degassing and cleaning applications generally employ unfocused acoustic radiation. In welding, soldering, drilling, machining and the neurosurgical controlled damage of brain tissue, on the other hand, a focused acoustic beam is always used.

Focusing may be done by the use of either an acoustic horn or by an acoustic lens. Figure 3

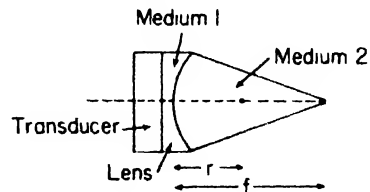


FIG. 3.

shows the principle of a simple plano-concave lens whose radius of curvature r is given by

$$r = f \left(\frac{n-1}{n} \right)$$

where n , the index of refraction, is the ratio of the velocities in the lens and adjacent medium:

$$n = \frac{V_1}{V_2}$$

Fundamental research in physics by ultrasonics. *Acoustic Velocity.* The acoustic velocity in a material is fundamentally related to the binding

forces between the atoms or molecules. Measurement of the velocity is made by a pulse technique or by vibrating a specimen of known dimensions in one of its fundamental modes of oscillation and recording the frequency.

Acoustic Attenuation. Attenuation measurements are made by recording the decay in oscillation amplitude with time of a solid material set to ring in one of its natural modes. Attenuation may also be measured from the amplitude of successive echoes passing through the material.

Under the influence of thermal activation, impurity elements or defects may switch position in a crystal lattice. This they do in a definite RELAXATION time, which is related exponentially to the temperature and to the difference in energy of the defect in its two positions. Now if the switching of position of the defect also causes a difference in dimension of the specimen, then a maximum in attenuation is found when the period of an applied acoustic stress just matches the thermal relaxation time. A whole spectrum of relaxation attenuation maxima have been found in solids. Each is related to a definite internal process, each with its characteristic energy difference. Perhaps the best known of these is the relaxation maximum given by the diffusion of carbon atoms in iron from which the activation energy for carbon diffusion can be calculated (Snoek peaks).

Dislocations are line defects in crystalline material the movement of which is involved in all cases of plastic deformation. Acoustic waves vibrate the dislocations, and in general, the more freely the dislocation moves, the greater is the acoustic beam attenuated. Attenuation measurements thus give information on how well the dislocation is locked in position in the solid, i.e., on how strong the material is. Pinning down of the dislocations may be effected by alloying or by radiation with high-energy particles which cause internal damage. Measurement of acoustic attenuation is therefore encountered in such studies.

Electrons in metals also attenuate acoustic waves if the frequencies are sufficiently high (>5 Mc/sec) and the temperature sufficiently low ($<10^\circ\text{K}$) so that the acoustic wavelength is comparable with the mean free path of the electrons. Attenuation measurements give information on relaxation times of electrons, and hence their energies, and are therefore concerned in the study of Fermi surfaces in metals. Such studies are intimately related to the phenomena of electrical and heat conductivity.

The onset of superconductivity in some metals and alloys at low temperatures is characterized by a fall to zero electrical resistance and also by a sharp fall in high-frequency acoustic attenuation.

In nonmetallic crystalline materials, thermal vibrations or PHONONS interact with one another and limit heat conductivity. Measurements of high-frequency attenuation is now giving information on phonon-phonon interactions. Such information may lead to the elucidation of the spectrum of thermal vibrations in all materials above the absolute zero of temperature.

T. S. HUTCHISON

References

- Goldman, R. G., "Ultrasonic Technology," New York, Reinhold Publishing Corp., 1962. Gives techniques used in ultrasonics.
Hutchison, T. S., "Ultrasonic Absorption in Solids," *Science*, **132** (Sept. 9, 1960). Treats the physics of ultrasonic studies.
Dransfeld, K., *Sci. Am.*, **208**, No. 6, 6 (1963). Information on recent high-frequency studies.

Cross-references: CONDUCTIVITY; ELECTRICAL; FERMI SURFACE; FERROELECTRICITY; HEAT TRANSFER; MAGNETOSTRICTION; OSCILLOSCOPE; PHONONS; REFLECTION; REFRACTION; RELAXATION; SOLID-STATE PHYSICS; SUPERCONDUCTIVITY; SONAR; VIBRATION; WAVE MOTION.

ULTRAVIOLET RADIATION

Ultraviolet radiation comprises the region of the electromagnetic spectrum extending from the violet end of the visible, wavelength 4000\AA , to the beginning of X-RAYS, arbitrarily taken as 100\AA , a span of more than five octaves. The unit of wavelength generally used for the ultraviolet is $\text{\AA} = 10^{-8}$ cm; it is named after A. J. Ångström, Swedish pioneer in SPECTROSCOPY, who made in 1868 the first accurate measurements of the wavelengths of spectral lines. Although still occasionally heard, the term ultraviolet rays has become obsolete. Ultraviolet light, however, is an expression which is usually acceptable even though ultraviolet cannot ordinarily be seen by the eye.

The ultraviolet is subdivided into several parts: The near ultraviolet, 4000 to 3000\AA , present in sunlight, producing important biological effects, but not detectable by the eye; the middle ultraviolet, 3000 to 2000\AA , called by biologists the far ultraviolet, not present in sunlight as it reaches the earth's surface, but well transmitted through air; the long range 2000 to 100\AA known as the extreme ultraviolet, abbreviated XUV, since it connects the ultraviolet and x-rays. The XUV is known also as the vacuum ultraviolet, because it is not transmitted through air, and is sometimes called the far ultraviolet. The portion of the XUV from 2000 to 1350\AA is known as the Schumann region, after its discoverer.

The boundary, 100\AA , between XUV and x-rays is arbitrary; it is often preferred to make the distinction on the basis of the method of production and analysis. Radiation is called x-rays if produced by the classical x-ray tube in which bombardment by electrons removes inner-shell electrons from atoms of the target material, with radiation emitted when outer electrons fall back; x-rays are also generated by the BREMSSTRAHLUNG process when fast electrons are suddenly decelerated. In sparks, arcs, and electrical discharges through gases, the classical sources of ultraviolet, it is the outer electrons of either neutral atoms or ions that are excited; the radiation occurs when the outer electrons are recaptured and fall back to their ground states (see ELECTRICAL DISCHARGES IN GASES).

Ultraviolet radiation was discovered by Ritter in 1801; he found that silver chloride was blackened, as it is by visible light, if placed beyond the violet end of the sun's spectrum, where nothing can be seen. Stokes, in 1862, using a prism of quartz rather than glass, observed to a short-wavelength limit of 1830\AA radiation produced by a spark discharge between aluminum electrodes by using a fluorescent plate detector. The breakthrough further into the ultraviolet was made between 1885 and 1903 by Victor Schumann, an instrument maker and machine shop owner of Leipzig. Schumann realized that there are three reasons why ultraviolet of shorter wavelengths had not been detected: (1) air is opaque, (2) quartz prisms and lenses do not transmit, (3) ordinary photographic emulsions are not sensitive, because of absorption by gelatin. He overcame these difficulties by constructing the first vacuum spectrograph, using optics of crystal fluorite instead of quartz, and by making photographic plates with almost no gelatin, now known as Schumann plates. Theodore Lyman of Harvard University, soon passed the limit, 1300\AA , reached by Schumann, by constructing a vacuum spectrograph with a reflection-type concave diffraction grating instead of the fluorite prism and lenses. He discovered the Lyman series of hydrogen, the most fundamental spectral series of the simplest and most abundant element, with first line, Lyman-alpha, at 1215.67\AA and series limit at 911.7\AA . Further progress reaching 140\AA was made by R. A. Millikan and his students at the California Institute of Technology with the aid of a "hot" spark in vacuum. Finally Osgood, in 1927, closed the gap to x-rays by combining the hot spark and a grating used at grating incidence; in the same year, Dauvillier reached 121\AA in the XUV from the x-ray side, using an x-ray tube with a spectrometer utilizing a crystal of large lattice constant, made from a fatty acid.

Ultraviolet radiation is emitted by almost all light sources, to some extent. In general, the higher the temperature, or the more energetic the excitation, the shorter are the wavelengths. Tungsten lamps in quartz envelopes radiate in the ultraviolet in accordance with Planck's law, slightly modified by the emissivity function of tungsten. Because of its high temperature, 3800°K , the crater of an open carbon arc is an excellent source of ultraviolet extending to the air cutoff. Electrical discharges through gases produce intense ultraviolet emission, mainly in lines and bands. The most widely used is the quartz mercury arc; when the Hg pressure is allowed to rise to several atmospheres, the intensity becomes great, and the spectrum is a quasi continuum. For the shortest wavelengths, still higher temperatures are required, as produced by discharging a large capacitor at some 50 kV between metal electrodes in vacuum. This violent discharge vaporizes atoms from the electrodes, then strips off as many as 10 or 15 outer electrons. The emission line radiation emitted when these highly ionized atoms recapture electrons extends to very short wavelengths. Another source producing highly ionized

atoms and emission lines in the XUV is the magnetically compressed plasma such as produced by the devices known as zeta and theta pinch, which reaches a temperature of half a million degrees. Still another useful source is the synchrotron, in which electrons are accelerated in circular paths through several hundred MeV. The electromagnetic radiation produced by the great centripetal acceleration is a continuum whose peak lies well below 100\AA . For great XUV intensity, however, no man-made source has equaled the atomic bomb.

Solids, liquids, and gases usually transmit well in the near ultraviolet but always become opaque somewhere in the middle or extreme ultraviolet. Among solids, both crystal and fused quartz transmit to a short wavelength limit between 2000 and 1500\AA , depending on purity; CaF_2 (fluorite), to 1230\AA ; and LiF , the most transparent known solid, to 1050\AA . These materials are invaluable for constructing lenses and prisms for ultraviolet instruments.

Solids in the form of very thin films transmit to some extent throughout the XUV, but the thickness must be less than a few thousand Ångströms. Certain thin films are useful as optical filters; Al, for example transmits from 837\AA corresponding to its plasmon frequency, to its x-ray L-edge at 170\AA , and to a small extent as far as 150\AA . Films of plastics, such as collodion of 300\AA thickness, transmit fairly well below 500\AA ; they are useful as windows for low-pressure gas retention.

Gases vary greatly in their absorption characteristics. Molecular oxygen causes air to become opaque below about 1850\AA , because of absorption in the Schumann-Runge band system, followed by continuous absorption from 1750 to 1290\AA , and strong irregular absorption to shorter wavelengths. Molecular nitrogen, however, is relatively transparent all the way to 1000\AA . Hydrogen absorbs in the Lyman series lines, and in an ionization continuum beyond the series limit, 911.7\AA . Of all the gases, helium is the most transparent; absorption first takes place in the resonance lines, the longest lying at 584\AA , and in a continuum beyond the series limit, 504\AA . More complicated molecular gases, such as CO_2 , NO, and N_2O are rather opaque throughout most of the XUV. Water vapor is much like O_2 with absorption commencing below 1850\AA .

Reflection occurs for ultraviolet, just as for visible radiation. In general, the reflectance becomes less, as the wavelength decreases. Aluminum is the best reflector over much of the long-wavelength region; when properly prepared, by rapid evaporation of the pure material in an excellent vacuum, the reflectance is greater than 90 per cent to 2000\AA , and can be maintained to 80 per cent at 1200\AA by overcoating with a thin layer of MgF_2 to prevent growth of Al_2O_3 . Below 1000\AA , platinum is best with a reflectance of about 20 per cent at 600\AA , but only about 4 per cent at 300\AA .

Ultraviolet can be detected in a variety of ways, making use of effects produced when it is absorbed by matter, e.g., fluorescence with reradiation of

longer wavelengths; chemical reactions in solids; dissociation of gas molecules, with ensuing reactions; ionization of gases; emission of photoelectrons from surfaces of solids. In general, the shorter the wavelength, the more energetic is the reaction. This is in accordance with Einstein's law, $E = h\nu$, giving the energy of a photon in terms of Planck's constant, h , and its frequency ν , (velocity of light \div wavelength). Thus, ultraviolet photons range up to about 10,000 times more energetic than visible photons, and effects produced by them are much easier to observe.

The simplest way to detect and measure ultraviolet radiation is by making use of the fluorescence process, converting the ultraviolet into visible radiation which can be seen, or into near ultraviolet which is easily photographed or measured with conventional photomultipliers. Calcium tungstate, for example, is an excellent converter for the middle ultraviolet. Materials much used for the extreme ultraviolet are oil and sodium salicylate. The latter is especially valuable because its quantum efficiency of fluorescence is high and is nearly independent of wavelength. An ordinary photomultiplier, with its glass window coated with a layer of sodium salicylate becomes a sensitive radiometer for use throughout the entire ultraviolet.

Ordinary photographic emulsions containing gelatin as a binder are useful only to about 2500Å; they can be sensitized easily by overcoating with oil or sodium salicylate. Eastman Kodak spectroscopic-type plates and films are available with ultraviolet sensitization, produced by overcoating with a fluorescent lacquer. Nearly gelatin-free Schumann-type emulsions combine greater sensitivity and higher resolving power than fluorescence-sensitized emulsions; they are available as Eastman Kodak SWR and SC-5, but must be handled carefully to avoid damaging the delicate surface.

Ultraviolet radiation is easily detected and measured directly by means of photomultipliers if they are equipped with ultraviolet transmitting windows or are used without an envelope in a vacuum. Almost all materials emit photoelectrons to some extent when ultraviolet-irradiated, but in applications, the problem is usually to devise a photocathode surface that has extremely high efficiency in a certain spectral region and is insensitive in others. When constructed with ultraviolet-transmitting envelopes, the various visible-sensitive photocathodes respond well throughout the ultraviolet. For applications in the presence of sunlight, however, it is often necessary to use a "solar blind" surface, having negligible sensitivity longward of 2900Å; one of the best surfaces is RbTe, with quantum efficiency of $< 10^{-3}$ at $\lambda > 3000\text{Å}$ and ≈ 0.1 at $\lambda < 2600\text{Å}$.

For use at wavelengths shorter than 1300Å the best photocathodes are simple metal surfaces, such as tungsten, used directly in vacuum. Most metals exhibit a strong internal photoelectric effect at $\lambda < 1500\text{Å}$, which reaches a high value of about 15 per cent at 1000Å. This high work function results in low sensitivity to long-wavelength

stray light, and low noise. Metal photocathodes are available in electrostatically focused photomultipliers with photocathode and dynodes of Be-Cu or stainless steel, and magnetically focused strip-photomultipliers with tungsten photocathodes. The sensitivity is little changed by repeated exposure to air.

Ionization chambers and Geiger counters form another useful class of detector of XUV radiation. Knowledge of the ionization efficiency of the gas makes it possible to use them for measurement of absolute energy. One of the most useful is filled with NO, responding from the ionization limit, 1350Å, to the transmission limit of the LiF window, 1050Å. To shorter wavelengths, it is possible to use Geiger counters without a window, by maintaining a slight positive gas pressure inside the tube. When filled with a rare gas, they count all incident photons of wavelengths shorter than the ionization limit of gas, since the gas ionization efficiency is 100 per cent.

Ultraviolet radiation can, of course, be detected with a thermocouple, by direct conversion of its energy into heat. Because of low sensitivity, this fundamental absolute method of energy measurement is resorted to only when no other method is available, for example, in order to establish that the ionization efficiency of the rare gases is 100 per cent for radiation below their series limits.

Mankind's principal source of energy, the sun, emits strongly throughout the ultraviolet, but only the near ultraviolet reaches the surface of the earth. Wavelengths shorter than 2900Å are absorbed by a layer of ozone (O_3) with center at an altitude of about fifteen miles.

The principal action of solar ultraviolet is to produce sunburn, or erythema, but it is only the shortest wavelengths penetrating the ozone layer which have this action. Since the effective band is centered at 2967Å and extends to about 3100Å, sunlight becomes rapidly more effective in burning the skin when the subject goes to a high altitude and when the sun lies high in the sky. Excessive exposure is known to be a cause of skin cancer.

The normal eye does not sense the sun's ultraviolet, although it does have a small sensitivity below 4000Å. Young eyes transmit more than old, and do, indeed, detect ultraviolet from 4000Å to about 3130Å as a faint, bluish sensation, but sharp images are not formed without special corrective lenses, on account of the chromatic aberration of the eye's optical system. To shorter wavelengths, ultraviolet is absorbed by the cornea and causes fluorescence, which is seen as a general haze. Excessive exposure to wavelengths short of 3000Å, however, causes conjunctivitis, a painful burn of the cornea. For this reason, it is extremely important to wear glass goggles, when in the presence of intense sources of middle ultraviolet radiation. Similarly, on snowfields and glaciers, goggles must be worn to prevent the form of conjunctivitis known as snow blindness.

A new and unfortunate action of solar ultraviolet is the production of Los Angeles-type smog, containing molecules which irritate the eye

and causes damage to plants. The photochemical processes are complicated and are not as yet completely understood. It is well established, however, that the principal atmospheric contaminants involved are nitrogen oxide, nitrogen dioxide, and various organic molecules present in gasoline engine exhausts. The initial process appears to be the absorption of solar ultraviolet by NO_2 , which produces O and NO. The principal reactions, however, are the photolysis of mixtures of nitrogen oxides and hydrocarbons in air, caused by absorption of solar ultraviolet. The products of the reactions are ozone, aldehydes, acrolein, acetone, and peroxyacyl nitrates and nitrites (PAN); among these, formaldehyde, acrolein, and PAN are specific eye irritants. One principal phytotoxicant is ozone, which produces a mottling or bleaching of the upper surfaces of leaves; various others cause a bronzing or glazing of the underneath surfaces of leaves. Obviously, solar ultraviolet cannot be eliminated; photochemical smog relief must come from the chemist.

The principal present-day application of ultraviolet radiation is in increasing the efficiency of conversion of electrical energy into light. The fluorescent lamp utilizes a coating of a crystalline substance, such as manganese-activated zinc silicate, on the inside wall of a mercury arc lamp, to convert the middle- and near-ultraviolet radiation from the mercury vapor column into visible light and thus add to the visible-line spectrum emission of the mercury. The efficiency of fluorescent lamps may reach values 2.5 times greater than those of commonly used incandescent tungsten lamps.

Another widespread use of ultraviolet is in "black lighting," largely for the theater. By introducing an ultraviolet-transmitting, visible-opaque filter over a carbon arc projector, an intense ultraviolet beam can be projected onto the stage. There it causes different materials to glow brilliantly, with color determined by the particular dye molecule.

As a technical industrial tool, ultraviolet spectroscopic analysis has become of extreme importance. In the production of steel, for example, a sample can be analyzed in minutes, by introducing it into a source electrode, and analyzing the radiation with a multichannel spectrometer. Different exit slits select the strongest and most sensitive emission lines of the various elements present; with photomultipliers, their intensities are measured and converted at once to give the composition of the steel.

Among various biological and medical applica-

tions of middle-ultraviolet radiation, perhaps the most important are its uses to kill bacteria and fungi in hospitals and especially in operating rooms, to eliminate the hepatitis virus from blood plasma, to keep foods sterile, and to treat skin diseases. Rickets and certain other diseases, can be cured by exposure of the body to ultraviolet, which produces vitamin D. Similarly, vitamin D is produced in milk by ultraviolet irradiation.

The middle- and extreme-ultraviolet radiation from the sun, although not able to reach the earth's surface, nevertheless affects man and his activities through its powerful influence on the upper atmosphere. First knowledge of the sun's ultraviolet spectrum was obtained in 1946, when a spectrograph was flown in a V-2 rocket by the U.S. Naval Research Laboratory and the solar spectrum was recorded from 3000 to 2200 Å. In the two decades since this event, great progress has been made in studying, from rockets and orbiting vehicles, the true solar spectrum, the reactions produced in the atmospheric gases when the sun's short wavelengths are absorbed, and the physical processes in the solar atmosphere giving rise to these radiations. Grouped together broadly as a space science, they comprise several fields, such as aeronomy, solar physics solar-terrestrial relationships, and ionospheric research. In the years to come, it is certain that orbiting observatories will monitor the ultraviolet and x-ray emissions from the sun, just as the great ground-based observatories study in visible and near-ultraviolet light the phenomena taking place in its atmosphere.

As affecting our present-day way of life, perhaps the most important action of solar short-ultraviolet and soft x-rays is in producing the several ionospheric layers; acting as mirrors, they reflect radio waves and so make possible radio communication over great distances. Far more important for mankind, however, is the solar radiation in the Schumann region, 2000 to 1300 Å. It is upon this radiation that the human race relies for survival; absorption of these wavelengths by molecular oxygen in the high atmosphere gives rise to atomic oxygen, which then reacts with molecular oxygen to form ozone. It is this permanent layer of ozone which protects all forms of terrestrial life from the lethal effects of the sun's middle-ultraviolet radiation.

RICHARD TOUSEY

Cross-references: BREMSSTRAHLUNG, ELECTRICAL DISCHARGES IN GASES, PHOTOELECTRICITY, PHOTOMULTIPLIER, PLASMAS, SPECTROSCOPY, SYNCHROTRON, X-RAYS.

V

VACUUM TECHNIQUES

The term "vacuum," which strictly implies the unrealizable ideal of a space entirely devoid of matter, is used in a relative sense in vacuum technique to denote gas pressures below the normal atmospheric pressure of 760 torr (1 torr = 1 mm of mercury). The degree or quality of the vacuum attained is indicated by the total pressure of the residual gases in the vessel which is pumped. Table I shows the accepted terminology in denoting degrees of vacuum together with the pressure range concerned, the calculated molecular density (from the equation $p = nkT$ where p is the pressure, n is the molecular density, i.e., number of molecules per cubic centimeter, k is Boltzmann's constant, and T is the absolute temperature taken to be 293 K or 20 °C) and the mean free path λ from the approximate equation for air: $\lambda = 5/p$ cm, where p is the pressure in millitorr.

Vacuum Pumps. There are several types of vacuum pumps. The two most widely used are the mechanical rotary oil-sealed pump and the vapor pump. The former provides a medium vacuum and works relative to the atmosphere; the vapor pump, on the other hand, provides a high or very high vacuum and operates relative to a medium vacuum provided by a rotary pump, referred to as a backing pump in this connection. Thus, the most widely used high-vacuum system able to establish an ultimate pressure of about 10^{-6} torr or below consists of a vapor pump backed by a rotary pump.

Four or five patterns of rotary oil-sealed pump exist, but they have in common the fact that the volume between a rotor (or rotating plunger) and a stator is divided into two crescent-shaped sections which are isolated from one another as

regards the passage of gas. Further, they are furnished with an intake port and a discharge outlet valve to the atmosphere. On revolution of the rotor (speeds of 450 to 700 rpm are used) gas is swept from the intake port, compressed, and discharged to the atmosphere via the one-way outlet valve. The mechanism is immersed in a low-vapor-pressure oil for sealing and lubrication in a small pump; larger units have a separate oil reservoir and feed device. A spring-loaded vane type of rotary oil-sealed pump is shown in Fig. 1(a). A single-stage pump of this kind provides an ultimate pressure of about 10^{-2} torr; a two-stage one with two units in cascade will give an ultimate of about 10^{-4} torr. Rotary pumps with speeds from 20 to 20 000 liters/min are commercially available, the smallest being driven by an $\frac{1}{4}$ -hp motor, the largest requiring a 40-hp motor.

These pumps handle permanent gases efficiently. Condensable vapors, e.g., water vapor, are not satisfactorily pumped because they may liquefy during the compression part of the rotation. To prevent this, gas ballast is a common provision whereby air from the atmosphere is admitted to the pump through a simple, adjustable screw valve to the region between the rotor and stator just before the discharge outlet valve. The amount of extra air admitted is readily adjusted to provide a compressed gas-vapor mixture which opens the discharge valve before vapor condensation occurs. Gas ballasting will clearly increase significantly the ultimate pressure provided by the pump, but this is not important since the gas-ballast valve can be closed after initial pumping has removed most of the water vapor.

Vapor pumps are of two main types: vapor diffusion pumps and vapor ejector pumps. Both

TABLE I. DEGREES OF VACUUM AND PRESSURE RANGES

Degree or Quality of Vacuum	Pressure Range (torr)	Molecular Density, n (molecules/cm ³)	Mean free path, λ (cm)
Coarse or rough vacuum	760--1	2.69×10^{19} — 3.5×10^{16}	6.6×10^{-6} — 5×10^{-3}
Medium vacuum	1— 10^{-3}	3.5×10^{16} — 3.5×10^{13}	5×10^{-3} —5
High vacuum	10^{-3} — 10^{-7}	3.5×10^{13} — 3.5×10^9	5— 5×10^4
Very high vacuum	10^{-7} — 10^{-9}	3.5×10^9 — 3.5×10^7	5×10^4 — 5×10^6
Ultrahigh vacuum	$< 10^{-9}$	$< 3.5 \times 10^7$	$> 5 \times 10^6$

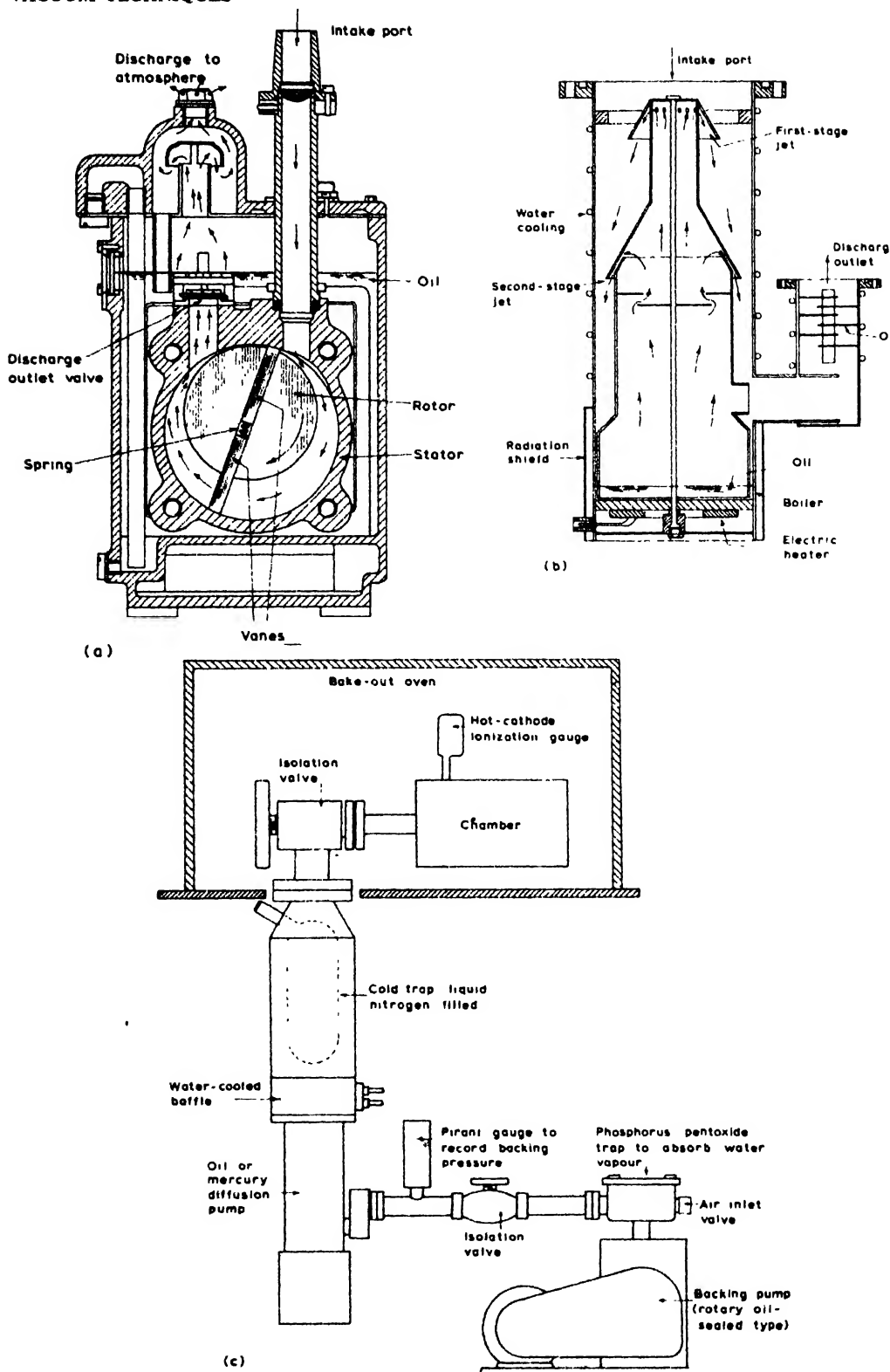


FIG. 1. (a) A spring-loaded vane type of rotary oil-sealed pump. (b) A two-stage oil diffusion pump. (c) A diffusion pump-rotary backing pump vacuum system.

employ vapor (of either mercury or a low-vapor-pressure oil) issuing from a jet as a means of driving gas in the direction from the intake port to the discharge outlet which is maintained at a medium vacuum by a backing rotary pump. In the diffusion pump [a two-stage design utilizing oil as the pump fluid is shown in Fig. 1(b)] the vapor issuing from the top, first-stage jet is directed downward towards the backing region. Gas molecules from the intake port diffuse into the streaming vapor. The directed oil molecules collide with the gas molecules to give them velocity components toward the backing region. A large pressure gradient is thereby established in the pump so that the intake pressure may be over 100 000 times less than the backing pressure. The intake pressure may therefore be 10^{-6} torr or lower with a backing pressure of 10^{-1} torr.

In the diffusion pump, the vapor stream is not essentially influenced by the gas pumped. In the vapor ejector pump, however, the vapor stream is enabled by a higher boiler pressure to be denser and of greater speed with a higher intake pressure, so that the gas is entrained by the high-speed vapor. Viscous drag and turbulent mixing now carry the gas at initially supersonic speeds down a pump housing of diminishing cross section. The ejector pump is designed to operate with a maximum pumping speed at an intake pressure of 10^{-1} to 10^{-3} torr and with a backing pressure of 0.5 to 1 torr or more. The diffusion pump, on the other hand, is designed to have a fairly constant speed from 10^{-3} torr down to an ultimate 10^{-9} torr or much lower in a modern, bakeable stainless steel system.

An important mechanical pump which operates in the same pressure region as the oil ejector is the Roots pump, capable of very great speeds and requiring backing by a rotary oil-sealed pump.

A vacuum system consisting of a diffusion pump and a backing pump, together with baffles, cold traps and isolation valves, is shown in Fig. 1(c). The cold trap is essential if a mercury vapor diffusion pump is used and is best filled with liquid nitrogen (-196°C); otherwise, the system will be exposed to the mercury vapor pressure, which is 10^{-3} torr at 18°C .

Ultrahigh-vacuum systems with stainless steel traps and metal sealing gaskets, and bakeable

(except for the pumps) for several hours to 450°C , may be constructed on the lines of that shown in Fig. 1(c) to provide an ultimate pressure of 10^{-9} to 10^{-10} torr.

Other vacuum pumps include the sorption type based on the high gas take-up of charcoal or molecular sieve material at liquid nitrogen temperatures. Sorption pumps may be used in place of rotary pumps, with a desirable freedom from rotary pump oil vapor, especially in systems where the amount of gas to be handled is limited.

The chief rival to the vapor diffusion pump at present is the getter-ion pump of the Penning discharge type, sometimes called the sputter-ion pump, with electrodes of titanium metal. The principle of operation is illustrated by Fig. 2 where an egg-box type anode is situated between plane cathodes. The anode-cathode operating potential difference is of the order of 2 to 10 kV, and the magnetic flux density is about 3000 gauss. The chief pumping action with active gases such as hydrogen, nitrogen and oxygen is to the anode which receives deposited titanium (which has very high gas affinity) sputtered from the cathodes under the action of the positive ion bombardment. Some gas, especially the inert gases like argon, is pumped to the cathodes.

A typical multicell pump of moderate size of this type has a pumping speed of about 250 liters/sec. Much larger pumps with speeds of up to 5000 liters/sec are commercially available, as are small single-cell units with speeds of some 2 liters/sec.

The sputter-ion pumps provide a vapor-free system giving a so-called dry vacuum, and they are often incorporated in plant with molecular sieve sorption as the backing pump. For medium-size laboratory plant able to provide ultrahigh vacuum they are most attractive. Probably their chief disadvantage is that the life of the pump is only about 40 hours at 10^{-3} torr, but this increases inversely with the pressure, so that it is 40 000 hours at 10^{-6} torr. At present, they are therefore strong rivals to the diffusion pump for plant where moderate amounts of gas are handled in the lower pressure ranges.

Getter-ion pumps of the type where the titanium metal is evaporated are also used, but the current

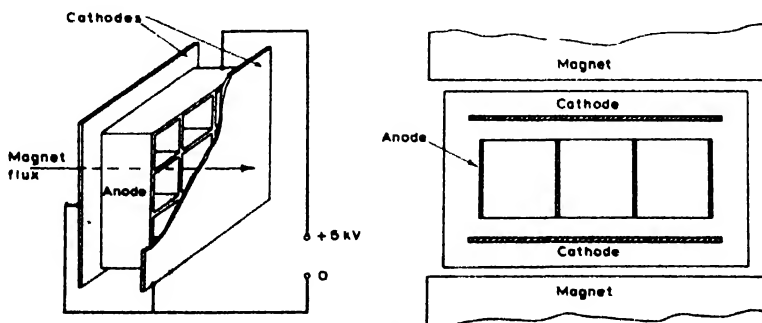


FIG. 2. A sputter-ion pump.

tendency is for the smaller sizes in glass to be popular, while the larger metal types are replaced by the sputter-ion variety.

Cryogenic pumping is presently receiving much attention. Here basically the provision is, within an initially evacuated system, of a surface which is at such a low temperature that gas impinging on it is condensed. For example, if the surface is maintained at the temperature of liquid helium (-269°C), all other gases have insignificantly low vapor pressures at this temperature and molecules of these gases impinging on the surface would remain there. A pumping speed for nitrogen of nearly $12 \text{ liters sec}^{-1} \text{ cm}^{-2}$ of cooled surface is hence theoretically possible. Liquid nitrogen, together with molecular sieve and other sorbent surfaces, and also liquid-hydrogen (-253°C , at which the vapor pressure of solid nitrogen is 10^{-10} torr) and liquid-helium cooled metallic surfaces are being actively investigated with the possibility of providing very high pumping speeds (10^6 liters/sec is not out of question) in space simulators and other plant.

Vacuum Gauges. A considerable problem in vacuum technique is that there is only one straightforward gauge able to measure low gas pressures absolutely in the sense that its calibration is independent of the nature of the gas and can be directly referred to millimeters of mercury: the McLeod gauge (Fig. 3).

The McLeod gauge is a compression device, i.e., the gas is compressed from an initial bulb of volume V into a small diameter capillary tube of volume v per unit length so that it occupies a length h of this capillary. With the setting indicated in Fig. 3, the pressure p is given by the equation

$$p = \frac{v}{V} h^2$$

where p is in torr if h is in millimeters and with v and V in the same units.

This gauge is clumsy, contains mercury and does not record correctly the partial pressure due to any condensable vapors present. Indeed, it is best to avoid it on most vacuum systems, but it is a virtually indispensable reference gauge for calibration work.

There are several alternatives, but none of them are absolute gauges and their calibrations all depend on the nature of the gas. Of the many possibilities, the thermal conductivity gauges of the Pirani and thermocouple types are useful within the pressure range from 10^{-4} to 10 torr, and operate by virtue of the dependence of the thermal conductivity of a gas on the pressure at low pressures. Two types of ionization gauge are valuable below 10^{-3} torr. The first of these is the Penning cold-cathode gauge with a range from 5×10^{-3} to 10^{-7} torr which has been extended in the inverted magnetron type (Redhead gauge) to 10^{-11} torr or below. The second is the hot-cathode ionization gauge of which the most widely used pattern is the Bayard-Alpert gauge (Fig. 4) which has become almost indispensable as a measuring

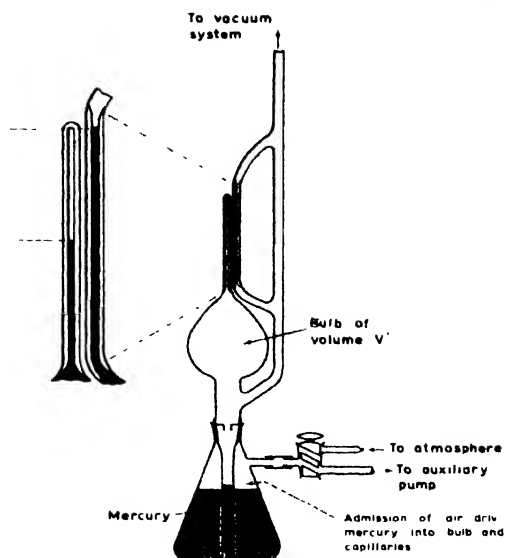


FIG. 3. The McLeod gauge.

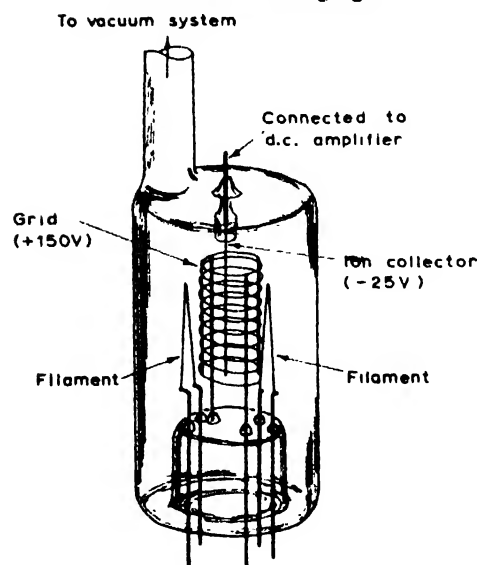


FIG. 4. The Bayard-Alpert hot-cathode ionization gauge.

instrument on ultrahigh-vacuum systems and has a range from 10^{-3} to 5×10^{-11} torr.

Within the hot-cathode ionization gauge, positive ions are created by impact with the residual gas molecules of the thermally emitted electrons in their paths to the positive grid (at about +150 volts w.r.t. the thermionic filament). These positive ions are attracted to the negative ion collector (at -25 to -50 volts) where the positive ion current as recorded by a calibrated dc amplifier is directly proportional to the gas pressure for a given gas and electron current. In order to minimize electron emission from the collector (indistinguishable from the arrival of positive ions) brought about by its irradiation with x-rays produced on

arrival of electrons at the positive grid, this collector is made of a thin central wire of insignificant interception of x-rays in the Bayard-Alpert gauge.

Such a gauge will typically have a sensitivity for nitrogen of about $20 \mu\text{A}$ per mA per millitorr, i.e., the positive ion current will be $20 \mu\text{A}$ with an electron current of 1 mA at a nitrogen pressure of 10^{-3} torr and it will decrease in direct proportion to the pressure.

These ionization gauges of both the hot- and cold-cathode varieties have a pumping action due to removal of positive ions to the collector and to the glass walls. This pumping action has been utilized in small glass systems to create an ultra-high vacuum. Initially, a pressure of 10^{-7} torr or below is established by a diffusion pump in the thoroughly baked ionization gauge. The gauge is then isolated from the pumps by a bakeable metal or greaseless glass valve. Subsequent operation of the gauge then reduces the pressure to the order of 10^{-10} torr. This technique, introduced by Alpert in 1950, has now been replaced by plant employing mercury or oil diffusion pumps with bakeable cold traps and metal gasket sealing, by the use of sputter-ion pumps and by cryopumping.

In the measurement of gas pressure, the determination of the partial pressures of the constituent gases is often as important as a knowledge of the total pressure. Gas analyzers for this purpose are based on the mass spectrometer of the magnetic deflection type, on the radio-frequency mass spectrometer, and on the omegatron. These gas analyzers also play an important part in leak detection techniques.

The applications of vacuum techniques are numerous and include the vacuum coating of substrates with metallic and insulating films in the production of optical mirrors, electrical resistors and capacitors, microminiature solid-state circuits, antireflection and enhanced reflection coatings on glass, conducting glass, interference filters, sorption and chemically reactive layers, etc. Further important fields of activity are the electron tube industry, vacuum drying and freeze drying, vacuum impregnation, distillation and molecular distillation, vacuum metallurgy including metal degassing and space simulation.

J. YARWOOD
K. J. CLOSE

References

- Dushman, S., "Scientific Foundations of Vacuum Technique," New York, John Wiley & Sons, Inc., 1962; second edition revised by Lafferty, J. M., Ed. Guthrie, A., "Vacuum Technology," New York, John Wiley & Sons, Inc., 1963.
Pirani, M., and Yarwood, J., "Principles of Vacuum Engineering," London, Chapman and Hall Ltd., 1962.

Cross-references: DIFFUSION IN FLUIDS, IONIZATION, KINETIC THEORY.

VACUUM TUBES. See ELECTRON TUBES.

VAPOR PRESSURE AND EVAPORATION

Vapor Pressure. Vapor pressure is the term applied to the driving force behind the apparently universal tendency for liquids and solids to disperse into the gaseous phase. All known liquids and solids possess this fundamental property, although in some cases it is too minute to be measurable. A typical liquid will exert a vapor pressure which is constant and reproducible. This pressure is dependent only upon the temperature of the system, and increases with increasing temperatures.

The molecular theory explains the phenomenon of vapor pressure through molecular activity. The molecules of a liquid are in rapid motion, even though they are in contact with each other. This motion or activity increases with temperature. At the vapor-liquid interface, this motion results in diffusion of some molecules from the liquid into the vapor. The attraction between molecules is strong, and some of the molecules dispersed into the vapor return to the liquid. The net number of molecules escaping produces the vapor pressure. For all practical purposes, this vapor pressure can be assumed constant whether the system is at equilibrium or not, due to the extremely high rate of molecular diffusion at the interface of the two phases.

In solids, the attractive forces of the molecules are so dominant that each is more or less frozen in place. Some diffusion does occur, however, as evidenced by the evaporation of ice, the odor of moth balls, and the slow diffusion or alloy formation of some metals kept in intimate contact. This vapor pressure increases with temperature, but is also a function of the molecular arrangement of the solid. As some solids such as sulfur are heated and molecular rearrangements take place, forming another allotrope of the same element, the vapor pressure changes sharply as this rearrangement occurs.

Vapor pressure can only be exhibited when the molecular activity is at a low enough level to permit continuous contact of the molecules and thus formation of a liquid. The maximum temperature at which this is possible is a fundamental property and is called the *critical temperature*. Above this temperature the material cannot be compressed to form a liquid, and only one phase results. This temperature is 705.4°F for water and -399.8°F for hydrogen.

The fundamental relationship between temperature and vapor pressure can be derived from thermodynamic laws. With certain limiting assumptions, the Clausius-Clapeyron equation is most often applied:

$$\frac{dp}{dT} = \frac{qp}{RT^2}$$

where q is the heat of vaporization. Because of these limiting assumptions, the integrated form of this equation is used in practice primarily as a guide to develop methods of correlating and plotting vapor pressure data.

The vapor pressure of a solution containing a nonvolatile substance (e.g., salt in water) is lower than that of the pure liquid. This phenomenon can again be explained by interference with the liquid molecular activity by the dissolved substances. The relationship between this vapor pressure depression and the concentration of the dissolved substance is valid for most substances at low concentrations. It was found to be dependent on the relative numbers of molecules of the solute and the solvent, and allowed accurate determinations of molecular weights of unknown solutes. If the Clausius-Clapeyron equation given above is combined with the above concentration relationship, it can be shown that:

$$\Delta T = \frac{RT^2}{q} \cdot C$$

where ΔT is the elevation of the boiling point, and C is the ratio of solute to solvent. This defines the effect of any solute on the vapor pressure exhibited by any solvent of latent heat q .

In the same manner, the vapor pressure of one component of a solution of two liquids has a different relationship with temperature than if it were pure. For many liquid mixtures, such as most hydrocarbon mixtures, the vapor pressure of the components vary directly from that exhibited in the pure form as their molar concentration in the solution. This relationship is known as Raoult's law:

$$\text{Partial pressure} = P_0 x$$

where x is the molar concentration of the component in the liquid and P_0 is the vapor pressure of the pure component at the same temperature as the mixture. A mixture following this rule is called an *ideal* solution, and its total volume is the sum of its components' volume.

If the gas phase above the liquid is also "ideal," the partial pressure of a component in this phase is equal to the total system pressure times the mole fraction of the component in the gas phase. This is called Dalton's law:

$$\text{Partial pressure} = P_t y$$

Combining these two formulas, it can be seen that:

$$\frac{y}{x} = \frac{P_0}{P_t} = K$$

for any particular temperature. This relationship can largely define many very complex liquid mixtures if the pressures used in the correlation are corrected by experimental data for deviation from the ideal.

A different relationship results if two liquids are relatively immiscible in each other. Molecular interference is minimal, and the total pressure exerted is equal to the sum of that of the individual pure components. The fundamental property of vapor pressure is thus dependent on temperature and composition of the material considered. These known and reproducible relationships have great technical application.

Vapor pressure relations can be used to determine heats of solution, heats of sublimation, and heats of fusion. Problems dealing with the solution of gases in liquids and adsorption of gases by solids are best handled by vapor pressure concepts. In dealing with solutions of miscible liquids, the most simple and useful relationship involves plotting the mole fraction y of one component in the vapor against x , the mole fraction of the component in the liquid. The ratio of $\frac{y}{x}$ is called

the phase equilibrium constant K and is used for definition of bubble points and dew points of simple and complex hydrocarbon mixtures over temperature and pressure ranges to near the critical.

Evaporation. The above effects of mixtures and of solutes on vapor pressure of a component are the very reasons why continuous generation of vapor (or *evaporation*) is the major tool of most process separations.

These effects permit us to separate salts from solutions, to separate liquid components from mixtures, and to use the energy relationships in process control of all kinds.

When molecules of a liquid do leave the surface and become vapor, they do so by overcoming the rather large attractive forces existing when they were in the liquid state. These forces were large since the molecules were in very close proximity in the liquid. Overcoming these attractive forces requires *energy*, heat energy, this is named the "*latent heat*" or "heat of vaporization" of the fluid. In general, this is in terms of heat units per unit of material, such as Btu per pound. In magnitude, latent heat will decrease as the liquid temperature increases or as the kinetic energy of the molecules increases. At the critical temperature, there is no latent heat—the molecules are in such an excited state that formation of a liquid is not possible.

For continuous evaporation, a continuous supply of energy is required. An available utility such as steam is a typical source.

Depending on the process, the evaporation is done:

(1) In equipment named "evaporators" such as the popular LTV of forced or natural circulation design used in acid concentration, salt production, sugar solution concentration and others.

(2) In processing distillation towers in a stepwise manner, tray to tray, resulting in a slight change in composition at each tray until required terminal conditions of overhead and bottom composition are reached.

(3) In reactors of various designs where one or more components are driven off frequently by the heat of reaction.

(4) In "cooling towers" where the desired effect is not the separation or concentration of components but the use of the latent heat of evaporating water to remove unwanted process heat.

(5) In all steam generating boilers.

(6) In any process step where a liquid/vapor phase change occurs.

The concentration and energy relationships discussed before apply.

Evaporation is thus the most widely used tool of nature and of industry. For this reason, an intimate knowledge of the theory of heat transfer and of the large quantity of empirical data available is essential to understand and develop almost any industrial process.

DOUGLAS L. ALLEN

Cross-references: GAS LAWS, HEAT TRANSFER, KINETIC THEORY, LIQUID STATE, MOLECULES AND MOLECULAR STRUCTURE, SOLID-STATE PHYSICS.

VECTOR PHYSICS

The Emergence of Vectors in Physics. A physical variable whose values can be specified by single numbers is called a *scalar concept*. The number 212 on the Fahrenheit thermometer specifies the boiling temperature of water at standard pressure. Temperature then is a scalar concept. By contrast, even when a suitable reference frame has been selected, the instantaneous velocity of a rocket is not fully described by stating that its speed is 2000 mph. If one also designates the *direction of the motion* (say by stating that it heads northeast and climbs at elevation 60°), then the velocity is unambiguously described. Velocity is an example of a *vector concept*.

Some of the special properties of vector concepts have been recognized for hundreds of years, e.g., the ancient fact that forces combine according to the familiar parallelogram law (see article on STATICS). Mathematical tools for dealing with directed quantities in three dimensions are outstanding products of nineteenth century mathematics. It now seems entirely natural to express key propositions of classical physics using vector algebra and vector calculus. A few decades ago, however, this was a controversial subject. Between 1891 and 1894, Peter G. Tait, Oliver Heaviside, Josiah Willard Gibbs, and others published fiery articles (still exciting reading) in the British journal *Nature*. Tait was a devoted promoter of quaternions, developed by his mentor, William Rowan Hamilton, as the proper tool for spatial physics. Gibbs, familiar with the work of Hamilton and also the more general theories of H. Grassmann, had devised his own treatment of physical problems. Heaviside was an indomitable advocate of the views of Gibbs. All of these men left their imprint on mathematics and physics. The notation of Gibbs, which mainly is followed here, has been particularly influential in the contemporary applications of vector mathematics.

Vector Algebra. To the question, "What is a vector?" one may give several reasonably respectable elementary answers. A vector is an arrow of a particular length and direction. A vector is a class of equivalent arrows. A vector is an ordered pair of points. A vector is a class of equivalent ordered pairs of points. A vector is an ordered triple of numbers. A vector is an ordered sequence of n numbers (which is called a scalar if $n = 1$!).

A vector is one of the undefined elements of a vector space. Each answer makes sense in its own context. The last answer provides a pattern which includes the others and offers the quickest approach to the algebra of vectors.

A *vector space* is an algebraic structure whose elements are called vectors (usually denoted here by boldface type: **A**, **B**, **C**, etc.), whose two operations are called *addition* and *multiplication by numbers* (we use real numbers throughout), and for which certain postulates are satisfied. These postulates should be regarded as a catalog of fundamental properties of such a structure. Other properties may be derived from them. The postulates may be stated as follows:

(1) For every pair of vectors **A**, **B** there is a vector sum such that

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

(2) For every triple of vectors, **A**, **B**, **C**, we have the equality

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$$

(3) There is a null vector **O**, such that for every vector **A**,

$$\mathbf{A} + \mathbf{O} = \mathbf{A}$$

(4) Every vector **A** has an opposite vector $-\mathbf{A}$, such that

$$\mathbf{A} + (-\mathbf{A}) = \mathbf{O}$$

(5) For every number c and every vector **A** there is a vector $c\mathbf{A}$ such that

$$c\mathbf{A} = \mathbf{A}c$$

(6) For every pair of numbers c , c' and every vector **A**,

$$(c + c')\mathbf{A} = c\mathbf{A} + c'\mathbf{A}, \quad c(c'\mathbf{A}) = (cc')\mathbf{A}$$

(7) For every pair of vectors **A**, **B** and every number c ,

$$c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$$

(8) The particular product $1\mathbf{A}$ is equal to **A**.

The geometric and physical significance of this list of algebraic properties is more clear when one considers suitable definitions of the two operations for particular vector spaces. For example, if the vectors represent forces, and hence are drawn as arrows, addition is composition of forces by the parallelogram rule, while multiplication by a number is merely an alteration of the magnitude (denoted by vertical bars) of the force according to the rule: $|c\mathbf{F}| = |c||\mathbf{F}|$. Multiplication by a negative number reverses direction. As a second example, if the vectors are triples of numbers (as in statics when components are listed), the two operations are at once defined thus:

$$(a_1, a_2, a_3) + (b_1, b_2, b_3)$$

$$= (a_1 + b_1, a_2 + b_2, a_3 + b_3)$$

$$c(a_1, a_2, a_3) = (ca_1, ca_2, ca_3)$$

One can easily verify that triples satisfy all eight postulates.

Two additional operations are of particular importance in physics. First, an *inner product* associates with each pair of vectors \mathbf{A}, \mathbf{B} a number, designated by $\mathbf{A} \cdot \mathbf{B}$, such that for all vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and every number c , the following identities hold:

$$(9) \mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$$

$$(10) \mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \cdot \mathbf{C}$$

$$(11) c(\mathbf{A} \cdot \mathbf{B}) = (c\mathbf{A}) \cdot \mathbf{B}$$

A final postulate for a vector space with inner product (called a *euclidean vector space*) is

$$(12) \mathbf{A} \cdot \mathbf{A} = 0 \text{ if and only if } \mathbf{A} = \mathbf{O}, \text{ and}$$

$$\mathbf{A} \cdot \mathbf{A} > 0 \text{ if and only if } \mathbf{A} \neq \mathbf{O}.$$

The physical concept of work provides a strong motivation for the usual definition of inner product. A force \mathbf{F} and a displacement \mathbf{S} are assigned a work $|\mathbf{F}||\mathbf{S}|\cos(\mathbf{F}, \mathbf{S})$ where (\mathbf{F}, \mathbf{S}) is the angle between vectors parallel to \mathbf{F} and \mathbf{S} . The defining formula for inner products of arrows is taken to be

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}||\mathbf{B}|\cos(\mathbf{A}, \mathbf{B})$$

Similarly, for triples

$$(a_1, a_2, a_3) \cdot (b_1, b_2, b_3) = a_1b_1 + a_2b_2 + a_3b_3$$

To prove that these two definitions satisfy postulates (9) through (12) is by no means trivial.

Similarly, an *outer product* associates with each ordered pair of vectors \mathbf{A}, \mathbf{B} a vector designated by $\mathbf{A} \times \mathbf{B}$, and possesses the following key properties. For all vectors $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and every number c ,

$$(13) \mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$$

$$(14) \mathbf{A} \times (\mathbf{B} + \mathbf{C}) = \mathbf{A} \times \mathbf{B} + \mathbf{A} \times \mathbf{C}$$

$$(15) (c\mathbf{A}) \times \mathbf{B} = c(\mathbf{A} \times \mathbf{B})$$

$$(16) \mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \cdot \mathbf{C})\mathbf{B} - (\mathbf{A} \cdot \mathbf{B})\mathbf{C}$$

The physical concept of moment of a force about a point leads naturally to a definition of outer product for arrows. Let \mathbf{F} be a vector representing a force acting at a point Q and for some point P let \mathbf{R} denote the arrow PQ . The moment \mathbf{M} of the force about the point P has magnitude given as follows:

$$|\mathbf{M}| = |\mathbf{R} \times \mathbf{F}| = |\mathbf{R}||\mathbf{F}|\sin(\mathbf{R}, \mathbf{F})$$

If $\sin(\mathbf{R}, \mathbf{F}) \neq 0$, the direction of \mathbf{M} is perpendicular to the plane of the point P and the arrow \mathbf{F} . We then write

$$\mathbf{M} = \mathbf{R} \times \mathbf{F} = [|\mathbf{R}||\mathbf{F}|\sin(\mathbf{R}, \mathbf{F})]\mathbf{N}$$

where \mathbf{N} is a vector of magnitude 1 (a *unit vector*) perpendicular to the PF plane. The sense of \mathbf{N} is usually taken so that the rotation of \mathbf{R} into the direction of \mathbf{F} determines \mathbf{N} by the right-hand rule. And for triples:

$$(a_1, a_2, a_3) \times (b_1, b_2, b_3)$$

$$= (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1)$$

For either of these models (arrows or triples), property (16) is quite remarkable. By applying it in two different ways to $\mathbf{A} \times [\mathbf{A} \times (\mathbf{A} \times \mathbf{B})]$, the reader may show, with the aid of property

(13), that $\mathbf{A} \cdot \mathbf{A} \times \mathbf{B}$ vanishes for all vectors \mathbf{A}, \mathbf{B} . Property (16) also leads easily to the Jacobi identity:

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) + \mathbf{B} \times (\mathbf{C} \times \mathbf{A}) + \mathbf{C} \times (\mathbf{A} \times \mathbf{B}) = \mathbf{O}$$

For the purposes of simple physical or geometric applications, the transition from arrows to triples is immediate. For a given reference frame, one lets (a_1, a_2, a_3) denote the three components (parallel to axes) of the arrow \mathbf{A} . These components may be computed by use of the inner product formula. Let $\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3$ be unit vectors parallel to three mutually perpendicular positive coordinate axes. Then, relative to these chosen axes, an arrow \mathbf{A} has components as follows:

$$a_1 = \mathbf{A} \cdot \mathbf{i}_1, a_2 = \mathbf{A} \cdot \mathbf{i}_2, a_3 = \mathbf{A} \cdot \mathbf{i}_3$$

Note that for each subscript j , since $|\mathbf{i}_j| = 1$, we have

$$a_j = |\mathbf{A}|\cos(\mathbf{A}, \mathbf{i}_j)$$

The preceding formula is used in statics for computing components of a force. One may think of each arrow \mathbf{A} as represented by the triple (a_1, a_2, a_3) of its components. In fact, one can make the merger complete by resolving each arrow into a sum of arrows parallel to the chosen axes:

$$\mathbf{A} = a_1\mathbf{i}_1 + a_2\mathbf{i}_2 + a_3\mathbf{i}_3$$

In this article, arrows and triples appear as alternate but related representations of physical concepts. For details on the ideas of this section see reference 2.

Vector Concepts in Elementary Geometry and Kinematics. In particle kinematics the most primitive ideas are position and displacement. Relative to an origin O the arrow OP is called the *position vector* of a particle at the point P . The displacement from P to P' is described by the arrow $\mathbf{D} = PP'$. One can easily see that $OP + PP' = OP'$, or $\mathbf{R} + \mathbf{D} = \mathbf{R}'$. It is natural now to introduce the notation of vector subtraction:

$$\mathbf{R}' = \mathbf{R} + \mathbf{R}' - \mathbf{R} = (-\mathbf{R}) + \mathbf{R}'$$

Hence, in vector geometry, a *displacement is equal to a difference of position vectors*. For a time interval Δt we often write the displacement as $\Delta \mathbf{R}$.

From the definitions of arrow operations, a number of geometric conclusions may be drawn:

(1) Non-null vectors \mathbf{A}, \mathbf{B} are *parallel* (or *anti-parallel*) if and only if $\mathbf{A} \times \mathbf{B} = \mathbf{O}$.

(2) Non-null vectors \mathbf{A}, \mathbf{B} are *perpendicular* if and only if $\mathbf{A} \cdot \mathbf{B} = 0$.

(3) An *equation for a plane* through a fixed point P' with position vector \mathbf{R}' and normal to a fixed vector \mathbf{N} is given by $(\mathbf{R} - \mathbf{R}') \cdot \mathbf{N} = 0$. This means for any point $P \neq P'$ on the plane, the arrow PP' is perpendicular to \mathbf{N} .

(4) Similarly, an *equation for a line* through P' parallel to \mathbf{N} is given by $(\mathbf{R} - \mathbf{R}') \times \mathbf{N} = \mathbf{O}$.

For further applications of vectors to elementary geometry, see Appendix 2 of reference 2.

Note that if one treats the position vector of P as a triple (x_1, x_2, x_3) , using coordinates, instead of as an arrow, the results just described are still valid.

If we define *velocity* as displacement per time, then we may represent the vector V as a derivative:

$$V = \lim_{\Delta t \rightarrow 0} (\Delta R / \Delta t) = dR/dt$$

Similarly, acceleration A is defined by $A = dV/dt$. Here again, when a reference frame has been chosen, one may employ triples:

$$V = (dx_1/dt, dx_2/dt, dx_3/dt)$$

$$A = (d^2x_1/dt^2, d^2x_2/dt^2, d^2x_3/dt^2)$$

In numerous physical applications, the use of vectors for geometric descriptions is effective. Consider a gas escaping through a surface element of area Δa . Call the outward unit vector normal to the surface element N . The area element may be treated as a vector

$$\Delta A = (\Delta a)N$$

If the velocity of escape is uniformly V , the outward flux (in particles per second) is given by

$$\Delta(\text{flux}) = \rho V \cdot \Delta A$$

where ρ is the number of particles per unit volume. Formulas for efflux under more general conditions may be found in books on kinetic theory, or in reference 2.

For a rigid body spinning about a fixed axis, the angular velocity is a vector ω directed along the axis of rotation according to the right-hand rule. Its time derivative is the angular acceleration α . If we take the origin of position vectors on the axis, the induced velocity and acceleration for a point of position vector R are given by

$$V = \omega \times R$$

$$A = \alpha \times R + \omega \times (\omega \times R)$$

The second term in the acceleration formula is called *centripetal acceleration*. Note that when $\alpha = 0$, $A = \omega \times V$ (see article on ROTATION — CIRCULAR MOTION). The preceding formula for centripetal acceleration in uniform circular motion has an interesting rotational analogue:

$$\alpha = \omega_p \times \omega_s$$

The subscripts stand, respectively, for precession and spin. This new formula describes uniform precession for gyroscopic motion (See article on GYROSCOPE).

Vector Concepts in Dynamics. Newton's second law (see article on DYNAMICS) may be summarized in vector equations:

$$F_1 + F_2 + \dots + F_n = (d/dt)(mV) = mA$$

The addition in the left member is vector addition. The product in the right member is multiplication

of a vector by a number. The left member may be written as

$$\sum_1^n F_i = \bar{F}$$

where \bar{F} denotes a resultant force. The abbreviated equation of motion

$$\bar{F} = m\bar{A}$$

where \bar{F} and \bar{A} stand for arrows, provides an expressive condensation of a basic physical proposition. The same equation serves as the key to analytical procedures if triples are used instead of arrows:

$$(\bar{f}_1, \bar{f}_2, \bar{f}_3) = m(d^2x_1/dt^2, d^2x_2/dt^2, d^2x_3/dt^2)$$

This vector equation yields three component equations:

$$\bar{f}_i = m d^2x_i/dt^2, \quad i = 1, 2, 3$$

In vector physics, the arrow representation of vector concepts repeatedly provides an easily visualized theoretical development, while the triple representation, for well-chosen axes, provides the strongest approach to computation. Some mechanical concepts are next listed in both forms.

Momentum of a particle:

$$mV = m(dx_1/dt, dx_2/dt, dx_3/dt)$$

Impulse exerted by a force:

$$\int_{t_0}^{t_1} F dt = \left(\int_{t_0}^{t_1} f_1 dt, \int_{t_0}^{t_1} f_2 dt, \int_{t_0}^{t_1} f_3 dt \right)$$

Work done by a force:

$$\int_R F \cdot dR = \int_{x_1}^{x_1'} f_1 dx_1 + \int_{x_2}^{x_2'} f_2 dx_2 + \int_{x_3}^{x_3'} f_3 dx_3$$

The integrands of the preceding equation must, of course, assume the values dictated by the path along which the force is allowed to act. For rotation about an axis, angular momentum is given by $I\omega$, I being the moment of inertia. In more general situations, the equivalent notion of moment of momentum is appropriate: $H = \int R \times V dm$. The details are not discussed here,

but, corresponding to $\sum F = \frac{d}{dt}(mV)$ we may write equations of the form

$$\sum R \times F = \frac{d}{dt} H$$

In rotation about a fixed axis, the second equation takes the familiar form

$$\sum M = \frac{d}{dt}(I\omega) = I\alpha$$

where each M is a moment about the axis, interpreted as a vector along it.

Vector Fields and Their Uses. A scalar field (scalar point function) assigns to each point (and

hence to each position vector) in a suitable domain a particular number. For example, to each point, \mathbf{R} , in a room is assigned a number called temperature, $\theta(\mathbf{R})$, another number called density, $\rho(\mathbf{R})$, another number called height above sea level, $h(\mathbf{R})$. Similarly, to each point in a room may be assigned a vector called gravitational field intensity, $\mathbf{G}(\mathbf{R})$, another vector called velocity (e.g., of air currents), $\mathbf{V}(\mathbf{R})$, another vector called magnetic flux density, $\mathbf{B}(\mathbf{R})$. These exemplify vector fields. When such scalar or vector fields have derivatives, important new fields may be derived. The pressure function in an ideal fluid is a scalar field $p(\mathbf{R})$. A related vector field is the pressure gradient, $\text{grad } p(\mathbf{R})$. The *pressure gradient* is a function assigning to each point the direction and magnitude of the maximum directional derivative of p . This new vector function is often written ∇p . Pressure gradient is intimately related to buoyant force: the net buoyant force per volume at \mathbf{R} equals $-\nabla p(\mathbf{R})$. Thus an equation for hydrostatic equilibrium is

$$\rho(\mathbf{R})\mathbf{G}(\mathbf{R}) = -\nabla p(\mathbf{R})$$

In terms of ordinary coordinates, the vector ∇p may be represented as triple:

$$\nabla p = (\partial p / \partial x_1, \partial p / \partial x_2, \partial p / \partial x_3)$$

Many other examples of gradient fields occur in physics. For instance, if a field intensity is conservative, it equals minus the gradient of a suitable potential function.

One simple vector field assigns to each position vector \mathbf{R} the corresponding velocity vector, $\mathbf{V}(\mathbf{R})$. For a rigid body rotating about a fixed axis, taking the origin on the axis, a formula for the velocity field is

$$\mathbf{V}(\mathbf{R}) = \boldsymbol{\omega} \times \mathbf{R}$$

where $\boldsymbol{\omega}$ is the instantaneous angular velocity. For the most general motion of a rigid body the formula is almost as simple: for any fixed point \mathbf{R}' of the body,

$$\mathbf{V}(\mathbf{R}) = \mathbf{V}(\mathbf{R}') + \boldsymbol{\omega} \times (\mathbf{R} - \mathbf{R}')$$

For details see reference 2.

The velocity pattern for a fluid will usually be more complicated than the one just considered. Compressions or expansions might occur. This sort of tendency may, at each point, be evaluated by the net outward flux per volume, called the *divergence* of the velocity field, written $\text{div } \mathbf{V}(\mathbf{R})$. Since this flux is measured in terms of volume rather than number of particles (as on p. 753), $\rho(\mathbf{R})$ is 1, and we have

$$\text{div } \mathbf{V}(\mathbf{R}) = \lim_{\Delta v \rightarrow 0} \frac{1}{\Delta v} \int_{\Delta v} \mathbf{V} \cdot d\mathbf{A}$$

Δa is the area of the solid element of volume Δv for which the net efflux per volume is expressed. Each such solid element contains the point with position vector \mathbf{R} . Divergence yields a scalar

field for a given vector field. Using cartesian coordinate triples

$$\text{div}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = \partial v_1 / \partial x_1 + \partial v_2 / \partial x_2 + \partial v_3 / \partial x_3$$

Since the net efflux per volume of mass is merely the rate of decrease of density, one can write at once the *equation of continuity* of hydromechanics

$$\text{div}(\rho \mathbf{V}) = -\partial \rho / \partial t$$

The concept of divergence has particular use in the theory of electromagnetism. For free space,

$$\text{div } \mathbf{B} = 0$$

$$\text{div } \mathbf{E} = 4\pi \rho$$

where \mathbf{B} is the magnetic flux density and \mathbf{E} the electric field intensity. This time ρ stands for electric charge density.

For a rigid body, the rotational aspect of a velocity field has been characterized by the vector $\boldsymbol{\omega}$. For a more general differentiable velocity field, one considers a derived vector function proportional to the local angular velocity vector. Only a formal cartesian expression for this new field, $\text{curl } \mathbf{V}$, is given here. For more general treatments not depending on particular coordinates, see reference 7.

$$\text{Curl}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) = (\partial v_3 / \partial x_2 - \partial v_2 / \partial x_3,$$

$$\partial v_1 / \partial x_3 - \partial v_3 / \partial x_1,$$

$$\partial v_2 / \partial x_1 - \partial v_1 / \partial x_2)$$

The concept of curl plays a vital role in many parts of physics. A field intensity is conservative if and only if its curl vanishes. The interrelationships between electrical and magnetic fields are most concisely expressed in terms of curl (see Maxwell's equations in most books on electromagnetism).

The formal expressions for gradient, divergence, and curl suggest regarding

$$\nabla = \mathbf{i}_1 \partial / \partial x_1 + \mathbf{i}_2 \partial / \partial x_2 + \mathbf{i}_3 \partial / \partial x_3$$

as a vector operator. Here we have used $\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_3$ as unit vectors parallel to coordinate axes, i.e., as abbreviations for the triples $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$. With this convention we may write:

$$\text{grad } p = \nabla p$$

$$\text{div } \mathbf{V} = \nabla \cdot \mathbf{V}$$

$$\text{curl } \mathbf{V} = \nabla \times \mathbf{V}$$

For more information about formal uses of ∇ see Appendix 3 of reference 2 or 8.

Extensions of Elementary Vector Physics. In preceding sections, we considered several vector functions of position. The function $\mathbf{V}(\mathbf{R}) = \boldsymbol{\omega} \times \mathbf{R}$ is a *linear* function, and hence can be represented in matrix notation. Equations for changes of coordinates can be expressed as translations combined with linear transformations (usually orthogonal). Linear transformations also occur in

expressions for strain in the theory of elasticity. Further details on linear transformations and vector physics ^{1,2,4,6} and treatments ^{1,5,7,10} of vectors in a more advanced mathematical setting can be found in the literature.

In this article we have considered mainly three-dimensional vector spaces. For the use of four-dimensions in special relativity,⁶ for vector treatments of variational mechanics,^{3,4} and for higher dimensional vectors used in phase spaces,² see the references listed below.

DAN E. CHRISTIE

References

The subject of vector physics is so inclusive that a suitable reference list would be very long, including books on classical mechanics, vector and tensor analysis, linear algebra, etc. Longer bibliographies will be found in references 2 and 4 below. Here we list only new or particularly pertinent sources.

1. Aris, R., "Vectors, Tensors, and the Basic Equations of Fluid Mechanics," Englewood Cliffs, N.J., Prentice-Hall, 1962.
2. Christie, D. E., "Vector Mechanics," Second edition, New York, The McGraw-Hill Book Co., 1964.
3. Coe, C. J., "Theoretical Mechanics, A Vectorial Treatment," New York, The Macmillan Co., 1938.
4. Eisenman, R. L., "Matrix Vector Analysis" (paperbound), New York, The McGraw-Hill Book Co., 1963.
5. Flanders, H., "Differential Forms with Applications to the Physical Sciences," New York, Academic Press, 1963.
6. Goldstein, H., "Classical Mechanics," Reading, Mass., Addison-Wesley, 1950.
7. Nickerson, H. K., Spencer, D. C., and Steenrod, N. E., "Advanced Calculus" (paperbound), Princeton, N.J., D. Van Nostrand, 1959.
8. Spiegel, M. R., "Theory and Problems of Vector Analysis," Schaum, 1959.
9. Schwartz, M., Green, S., and Rutledge, W. A., "Vector Analysis with Applications to Geometry and Physics," New York, Harper and Row, 1960.
10. Wrede, R. C., "Vector and Tensor Analysis," New York, John Wiley & Sons, 1963.

Cross-references: ELASTICITY, ELECTROMAGNETIC THEORY, FIELD THEORY, FLUID DYNAMICS, GYROSCOPE, MECHANICS, RELATIVITY, ROTATION—CIRCULAR MOTION, STATICS.

VELOCITY OF LIGHT

Every elementary electric charge occupies a small volume. The charge is surrounded by a radially directed electric field, the strength of which decreases by the square of distance. By any motion with a velocity $v > 0$ of the charge relative to some fixed point, the distance and/or direction is changed and, thereby, the (vector) value of the field strength at the point. Here the field change occurs a while after the moment of charge motion. Otherwise expressed, the field change propagates with a finite velocity, usually labeled c . Moreover,

during the short time for the action from a *certain part* of the charge volume to pass the rest of it, the volume moves a small distance proportional to v/c . This very small displacement creates the magnetic field always associated with an electric field change. In fact, magnetism is due to a, usually very small, part of the electric field.

Any change in the electromagnetic field propagates with the very high velocity of $c = 3 \times 10^{10}$ cm/sec, called the velocity of light. For, if the charge motion happens to be an oscillation of frequency ν between 4×10^{14} and 8×10^{14} cps, we note the corresponding field variations as visible light. According to definition it is

$$c = \nu \cdot \lambda \quad (1)$$

where λ is the wavelength in vacuum. The movements of the elementary charges are mostly oscillatory. By means of the field, after the delay due to c , they act on surrounding charges, and in that way all events in the atomistic world depend on the value of c . Therefore, the knowledge of c has turned out to be of extreme importance to our modern civilization.

The wave velocity according to Eq. (1) has a constant value, independent of all movements and, strictly speaking, independent of the medium where the propagation takes place. In a transparent body, in the intermediate space the action among the elementary charges of the atoms disperses with the velocity c . By inertia, the oscillating of the charges due to the active field is somewhat delayed as compared to the field. Now, the oscillating charge itself is a radiating source and its own delayed field interferes with the original one, creating a sum-wave. The velocity v of which is $v < c$. Exceptionally, in cases of resonance and absorption $v > c$. Our observation of light propagation in a substantial body relates to that slower interference wave. For the body, e.g., the medium of atmospheric air, we obtain corresponding to Eq. (1):

$$v = \nu \cdot \lambda_a \quad (2)$$

From plain optical geometry we know that the refractive index of a medium is

$$n = \frac{c}{v} = \frac{\lambda}{\lambda_a} \quad (3)$$

The technique for determining the wavelengths of light has progressed very far, and one can easily obtain an accurate value of n . For visible light, n depends on the wavelength used, according to

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \quad (4)$$

where A , B and C are positive constants. Close to resonance, the formula is not valid.

Direct determinations of the velocity of light, usually performed in air, are all based on the measurement of the time for a light pulse to cover a known distance. Such a pulse means an increase, followed by a decrease of the amplitude of the light vibrations. What we observe is energy exchange associated with the amplitude changes,

and as a result, we obtain the propagation velocity of the light energy. A change of amplitude is, however, equivalent to an interference among a series of adjacent wavelengths, since that change is created by just such an interference. The light pulse, therefore, consists of a whole group of adjacent wavelengths, interfering with each other. Interference is a sum-product. If the participating waves have different velocities, one will find by simple addition of two sine oscillations, that the group formed has a velocity different from those of the waves creating the group. On the surface of a calm sea, we observe a group of waves moving forward. The waves are created at the back of the group, travel through it, and disappear at the front. The difference between wave and group velocity is directly proportional to the wavelength and to the dependence of the wave velocity on the wavelength. Thus, calling the group velocity u , we obtain the difference:

$$v - u = \lambda_n \cdot \frac{dv}{d\lambda_n} \quad (5)$$

Analogously to Eq. (3), we introduce a "group index"

$$n_g = \frac{c}{u} \quad (6)$$

If, as in the case of air, $dv/d\lambda_n$ is a small quantity, we get, after some recalculation,

$$n_g = n - \lambda \cdot \frac{dn}{d\lambda} \quad (7)$$

From Eq. (4) it is evident that $dn/d\lambda$ is a negative quantity, i.e., $n_g > n$. In vacuum, all velocities are equal: $u = v = c$.

The group velocity refers to the energy transport. Thus, in a medium there are the original waves of vacuum velocity c . By interference with waves from local charge oscillators, they create a wave system characteristic of the medium being considered and with the velocity $v < c$. By external energy action, that system is divided into an increasing number of adjacent waves, again interfering, of "second order," thereby forming groups of velocity $u < v$ (if normally $dn/d\lambda < 0$). Application of Eq. (7) to Eq. (4) yields:

$$n_g = A + \frac{3B}{\lambda^2} + \frac{5C}{\lambda^4} \quad (8)$$

In the case of visible light or $\lambda = 0.4 - 0.8 \mu$ ($\mu = 0.001$ mm) and in dry air of 0°C , 760 mm Hg, $A = 1.000287619$, $B = 16.204 \times 10^{-7}$, $C = 0.1391 \times 10^{-7}$ (based on values derived by Edlen). Inserting n_{0g} corresponding to the λ used, we get

$$n_g = 1 + \frac{n_{0g} - 1}{1 + \alpha t} \cdot \frac{p}{760} - 0.55 \times 10^{-7} \cdot \frac{e}{1 + \alpha t} \quad (9)$$

where t is expressed in degrees Celsius, p in millimeters Hg, e in millimeters Hg of humidity, and $\alpha = 1/273$. For visible light, the error of Eq. (9) is less than 5×10^{-8} .

For micro- and radar waves with $\lambda > 700 \mu$ the influence of λ in Eq. (8) is insignificant. Increasing influence of humidity, however, makes

$$n = 1 + \left[\frac{103.64}{T} (p - e) + \frac{86.26}{T} \left(1 + \frac{5748}{T} \right) e \right] \times 10^{-6} \quad (10)$$

where $T = t + 273.16$. By using 103.64 in place of 103.49 as recommended by the International Association of Geodesy, the influence of a CO_2 -term is partly compensated for. The uncertainty of Eq. (10) is about 5×10^{-7} .

There are a variety of methods for the determination of c . Among the direct ones, is the measurement of the travel time for a light pulse to cover a known distance. Usually there is a continuous series of pulses; i.e., the intensity is varied with a definite and known frequency. At every moment, there is a definite state of phase of the variation period. After reflection, the light returns, and one makes a phase comparison between emitted and received intensity. Thus, the intensity variation is used as a clock for the measurement of the running time. If the phase comparison is based on having the maximum of emitted light intensity coincide with the maximum of received intensity, the running time evidently is an even multiple of the period (= time) of a complete intensity variation cycle. Usually only one definite multiple number yields reasonable values, and there is no need for a special determination of the multiple.

The indirect methods are of much more varied character.

Determination of c . Galileo was the first to try to show a finite light velocity. In his direct method, he used lanterns which could be screened off rapidly. The time elapsed on a distance of few kilometers was, of course, too small to be observed visually.

In August 1676 at Paris, the Danish astronomer Ole Römer determined the revolution period of a satellite of Jupiter by its eclipse times into the planet's shadow. He used the result, 42 hours, to calculate the point of time for an eclipse occurring in November. It really occurred 10 minutes later. Römer explained this by the longer time for the light to reach the earth, now at a considerably greater distance from Jupiter. From the dimensions of the earth's orbit Römer defined c to be

$$c = 214\,300 \text{ km/sec}$$

A modern value according to the same method is $299\,840 \pm 60$ km/sec.

In 1725 the Englishman Bradley discovered astronomic aberration. Due to the velocity v of the earth in her orbit, the apparent direction to the stars normal to v is changed by v/c . Bradley could detect and measure the angle v/c because of its alternate sign for opposite parts of the earth orbit. Knowing v Bradley computed

$$c = 295\,000 \text{ km/sec}$$

A modern value is $299\,857 \pm 120$ km/sec.

In 1849 the Frenchman Fizeau took the next step. By projection of the image of a point source on the edge of a rapidly revolving cogged wheel, he got the required light pulse series. A blink or pulse passing through a cog interspace traveled the distance of 8633 meters to a reflector and back again along the optically identical path. Once more, now in the opposite direction, the pulse reached the cogged edge. For a suitable rotational speed the next cog interspace had now moved into position to let through the pulse; a case of "coincident maxs" (maxima). The time was computed from the revolution speed and the number of cogs. Fizeau's value was:

$$c = 315\,300 \pm 500 \text{ km/sec}$$

In 1850, after an idea of Arago, using a point light source, Fizeau's compatriot Foucault directed a beam of constant intensity to a rapidly revolving plane mirror. During the time for the light to cover the distance to a second fixed mirror and back, the revolving mirror had turned through a small angle, and the image of the light source was slightly displaced from its position when the mirror was stationary. Foucault's value was:

$$c = 298\,000 \pm 500 \text{ km/sec}$$

In 1856 the German scientists Kohlrausch and Weber achieved a very important indirect determination. The magnetic field strength depends on rc , the charge velocity r creating the magnetic field. Therefore, based on the same unity of force, the ratio between the unit of charge in electric and in magnetic measuring systems becomes c . Kohlrausch and Weber measured a charge (1) by its attraction to a second equal charge and (2) by the current produced at the passage of the charge through a galvanometer which showed a deflection due to the magnetic force. From (1)/(2) c , they got

$$c = 310\,800 \text{ km/sec}$$

Rosa and Dorsey in 1906 measuring a capacity obtained $299\,784 \pm 30 \text{ km/sec}$.

In 1891 the Frenchman Blondlot transmitted Hertz's waves along two straight and parallel wires, a Lecher guide system. After reflection, the waves created a standing wave system. Observing the nodal points Blondlot determined the wavelength. The frequency was known, and using Eq. (1) he obtained:

$$295\,000 \pm c = 305\,000 \text{ km/sec}$$

At this time, however, one was very uncertain as to the ratio between the velocities in vacuum and along the wires. Gutton solved this problem in 1911 by aid of the Kerr cell, making possible light variations of a frequency equal to that applied to the Lecher system. In a Kerr cell, the light passes between two condenser plates, lowered into nitrobenzene. By the electric field between the plates, due to the directive action on the oblong dipole molecules, the fluid acquires optical double-refracting properties. As soon as

the field ceases, the molecules recover their random directions through the influence of their thermal movements. Placing the cell between an optic polarizer and a normally oriented analyzer, the light intensity leaving the combination, within certain limits, is directly proportional to the voltage applied to the condenser. The voltage may be that from a high-frequency radio transmitter.

Later on Gutton's compatriot Mercier derived a theoretical formula for the ratio between light and the velocities of guided waves. Using a valve oscillator for the high-frequency voltage on a Lecher system and applying his formula to the result, he obtained, in 1921.

$$c = 299\,782 \pm 30 \text{ km/sec}$$

In the 1870's, Newcomb had introduced a revolving reflecting multi-surface prism in place of Foucault's mirror. Michelson improved the system further, first in 1879 and finally at Mt. Wilson in 1926, his most accurate measurement. During the time required for the light to cover the distance of 35 km and back, the prism turned the next of its 12 surfaces in place for reflection. Only for exactly correct rotational speed the position of the image of the suddenly illuminated distant reflector in the field of sight was independent of the direction of rotation. The comparison of phase may be said, here too, to have occurred at "coincident maxs," and the high accuracy was due to the great displacement of the image for a turn of the revolving prism. Michelson's value was

$$c = 299\,796 \pm 4 \text{ km/sec}$$

He overlooked a group correction of $\pm 2 \text{ km/sec}$.

Michelson, Pease and Pearson planned a determination in vacuum. It was completed by Pease and Pearson after Michelson's death in 1931. By multiple reflection in a 1-mile evacuated pipe, the light path was 10 miles. Due to the pipe, the geometry of the measuring device was considerably less symmetrical than in the case of the Mt. Wilson determination. The experiment was much talked about and the result was considered of high quality.

$$c = 299\,774 \pm 11 \text{ km/sec}$$

In 1928 the German scientists Karolus and Mittelstaedt performed the first direct determination using a Kerr cell. Their result was

$$c = 299\,778 \pm 20 \text{ km/sec}$$

Here too the group correction seems to have been overlooked.

Now Michelson's 1926 result was regarded as doubtful, and in 1940 Anderson, an American, tried to arrive at a decision. He used a Kerr cell. For phase comparison a photomultiplier received light from two nearby mirrors, the second of them at a somewhat greater distance in order to obtain intensity maxima from the one mirror coincident with minima from the second. The alternating photo-current then was equal to zero. Thereupon, the second mirror was moved away

some 170 meters, and the distance was adjusted anew to yield a photo-current equal to zero. After the known mirror displacement, the light's travel time had increased an even multiple of the intensity period, known from the frequency. Anderson's value:

$$c = 299\,776 \pm 14 \text{ km/sec}$$

Thereby the Michelson 1926 value was ruled out. Anderson carefully considered the group correction. Possibly, Anderson on occasions of his apparatus yielding values very different from the evacuated pipe result, might have felt particularly called upon to search for error sources and in that way unconsciously influenced the result. Or it might have been pure chance. In any case, his value superceded the pipe result.

Anderson's value was used for radar systems. Thereby a microwave transmitter emits short pulses (10⁻⁶ second) which after distant reflection and return are received again. On the screen of a cathode ray tube the rapid sweep of the ray is marked by emitted and received pulses. The difference of mark positions and the sweep speed yields the pulse travel time. By knowing c the distance is obtained. "Oboc" (0.1-meter wave) and "Shoran" (1-meter) applied to known distances gave low values. Was the c -value used too low?

In 1947 Jones and Conford used Oboc in a manner opposite to its usual operation for the determination of c over known distances and got

$$c = 299\,782 \pm 25 \text{ km/sec}$$

In the case of Shoran the main pulse-giving radar station is in an aircraft traversing the known straight distance between two points. In order to get strong echoes, there are radar slave stations at the points. Immediately after receiving the pulses, the slave stations reemit them on a slightly different wavelength. By that the main station receiver is not disturbed by its own transmitter. The distances to the two slave stations are continuously registered, and the shortest distance and also that used are obtained at the moment of transversing. A typical distance is 300 km. By careful treatment of Shoran results, Aslakson in 1949 obtained a velocity value

$$c = 299\,792.3 \pm 2.3 \text{ km/sec}$$

In 1951 he used "Hiran," by which errors from varying signal strengths are avoided, and obtained

$$c = 299\,794.2 \pm 1 \text{ km/sec}$$

Meanwhile Essen attacked the problem in a quite new manner by his indirect determination using a cavity resonator. Microwave guides may be circular tubes of sufficient diameter to let the waves through. If such a tube, whose length is an even multiple of half the wavelength, is closed by plane reflecting covers, the oscillating energy supplied remains in the cavity as a standing wave. This state only occurs in case of an exact, definite resonance frequency. In practice, there are several adjacent resonance frequencies depending

on remaining oscillation states, e.g., including different multiples over a diameter or over several symmetrically orientated diameters. Sarbacher and Edson have derived the exact mathematical expression for all these resonance frequencies as functions of the tube length, tube diameter and the inside wave velocity. Essen and Gordon-Smith very carefully measured length, diameter and frequency and used an evacuated cavity. By calculations, they found the vacuum value c . Their result, in 1947, was:

$$c = 299\,792.2 \pm 4.5 \text{ km/sec}$$

In 1949 Essen repeated the determination using a cavity whose length could be varied by an inserted movable plunger. By the dependence of the results on the cavity length, systematic errors could be rejected. The final value was

$$c = 299\,792.5 \pm 1 \text{ km/sec}$$

Applying current practices, the error limits of 1 km/sec may be reduced perhaps by half.

At this time, Bergstrand performed direct experiments using visible light. The alternating voltage, supplied by a 10-Mc/sec radio transmitter, coincidentally fed Kerr cell and photo tube. In place of "coincident max's" the phase comparison was carried out in moments of light pulses reaching half of maximum intensity. Then the photo-current was strongly dependent on the distance to the reflector. By low-frequency phase shift of the emitted light intensity modulation (interchange of maximum and minimum), there were periodically two photocurrents, and by exactly chosen distance, they balanced each other to an O-current on the control instrument. In this way, by known displacement of the reflector through several successive such O-points the wavelength of intensity variation, i.e., the group length L_K , could be determined. The group velocity was obtained from $u = vL_K$, reduced to c by Eqs. (6) and (9). Besides some preliminary results from 1947-48, the 1950 value was:

$$c = 299\,793.1 \pm 0.3 \text{ km/sec}$$

Under the name of "Geodimeter," Bergstrand's instrument is used all over the world for measuring distances with high accuracy, c being considered as known. The error in 30 km need not exceed 5 cm. When the geodimeter was used to check the rocket-camera positions around Cape Kennedy, the accuracy turned out to be still better.

Using the geodimeter on known distances, of course, one can obtain c . The experience from recent, improved models now shows that the just related value of 1950 is slightly too high. In 1953 Mackenzie on the Ridge Way and Caithness base lines obtained

$$c = 299\,792.3 \pm 0.5 \text{ km/sec}$$

In 1959 a Russian geodimeter determination on 6 base lines yielded

$$c = 299\,792.5 \pm 0.1 \text{ km/sec}$$

Almost identical results were obtained on Czech and Bavarian base lines. In 1955 Velitschko constructed a Kerr apparatus (SWW-1) for distance measurements. There was a Kerr cell on the receiving end also. The phase indication was visual. On base lines, the SWW-1 yielded:

$$c = 299\,792.7 \pm 0.3 \text{ km/sec}$$

Intended for the same purpose, but less accurate than the geodimeter, are instruments using microwave (10 cm.) radiation sources, such as Wadley's Tellurometer of 1957. By this method, a moving wave system is created between two equal instruments, having slightly different measuring frequencies (10,000 and 10,001 Mc/sec). The phase comparison is done by the low-beat frequency of 1000 cps.

Generally, perhaps the hitherto most accurate c determination is that by Froome. In 1950 he started preliminary measurements and in 1958 he used a klystron transmitter delivering harmonics of 0.4 cm wavelength. By guiding pipes the microwave emission was divided into two equal parts. The two pipes terminated in transmitting horns supplied with lenses to form the front surface of the emitted wave as plane as possible. The pipes were bent 180° each in order to get the horns directed opposite to each other. The distance between the horns was 6 to 14 meters. Between the horns there was created a standing wave system, where the nodal points were observed by aid of a movable receiving detector. The displacement of the detector was measured interferometrically. In this way the microwave length was determined with great accuracy. By carrying through the determination for different horn distances, systematic errors were rejected. These errors may depend on disturbing reflections or on the wave fronts not being sufficiently plane or of known shape. The frequency was controlled by aid of an atomic clock. The result of 1958 was

$$c = 299\,792.5 \pm 0.1 \text{ km/sec}$$

or identical to the Russian and other recent geodimeter results. By use of a maser, Froome expects to realize an analogous measurement with visible or infrared light.

A summary of determinations of recent years thus indicates a c -value very close to 299 792.5 km/sec.

Finally, there is an indirect method which does not give such high accuracy: The exactly known frequency of an atomic clock is the ground rotational molecular frequency with quantum number $= 1$ of some diatomic gas. From the quantum spectral formula of the rotational line series, one obtains the frequency at high quantum numbers of visible or infrared light, where the wavelength is known with great accuracy. The connection between wavelength and frequency yields c , and in 1955 Rank, Bennet and Bennet obtained the value $299\,791.9 \pm 2.2 \text{ km/sec}$.

ERIK BERGSTRAND

References

- Brillouin, Léon, "Wave Propagation and Group Velocity," New York, Academic Press, 1960.
 Michelson, A., *Astrophys. J.*, **65** (1927).
 Michelson, A., *Astrophys. J.*, **82** (1935).
 Anderson, J. *Opt. Soc. Am.*, **31** (1941).
 Essen, *Proc. Roy. Soc. London Ser. A*, **204** (1950).
 Froome, *Proc. Roy. Soc. London Ser. A*, **109** (1958).
 Aslaksen, *Trans. Am. Geophys. Union*, **32** (1951).
 Dorsey, *Trans. Am. Phil. Soc.*, **34**, Pt. 1 (1944).
 Bergstrand, E., in Flugge S., Ed., "Encyclopedia of Physics," Vol. 24, Berlin, Springer, 1957.

Cross-references: ATOMIC CLOCKS; CONSTANTS, FUNDAMENTAL; ELECTROMAGNETIC THEORY; INTERFERENCE AND INTERFEROMETRY; KERR EFFECTS; OPTICS, GEOMETRICAL; PROPAGATION OF ELECTROMAGNETIC WAVES; RADAR; WAVEGUIDES; WAVE MOTION.

VIBRATION

In physics, the term "vibration" is used for any sustained motion, characteristic of a finite system, in which each particle or element of the system moves to-and-fro about an equilibrium position. In the simplest situation, such a motion possesses a unique "periodic time." The inverse of this quantity, the "frequency," is the number of complete to-and-fro excursions per unit time.

It might appear that reference to a finite system in the above definition is gratuitous: infinite systems are mere figments of the theorist's imagination; only finite systems have real existence. Nevertheless, the emphasis is intentional. As will appear, the natural frequencies of vibrating systems depend, in general, on the size (spatial extent) of such systems.

In order to introduce the subject, the motion of a simple pendulum will first be considered—though, strictly, this motion does not fulfil the conditions that have been specified for vibrations: the natural frequency is not characteristic of the pendulum alone, it depends also on the acceleration due to gravity at the place where the pendulum is used. It is usual to speak, therefore, of the oscillations of a pendulum, rather than of its vibrations.

An ideal simple pendulum consists of a very small massive bob (of mass m) attached to one end of a perfectly flexible massless string (of length l), the other end of the string being fixed to a rigid support. In equilibrium, the pendulum hangs vertically; when the bob is displaced sideways and let go, the pendulum oscillates in a vertical plane. In such motion, the string, being massless and under tension, remains straight, and the whole system may be regarded as rigid. An ideal simple pendulum, then, is effectively a rigid body of moment of inertia ml^2 about its axis of oscillation, and, the moment of the weight of the bob about this axis being $mg l \sin \theta$, when the angular displacement is θ , the equation of motion is

$$ml^2 \ddot{\theta} = -mg l \sin \theta$$

Here g is the acceleration due to gravity. For small displacements, therefore,

$$\ddot{\theta} = -(g/l)\theta \quad (1)$$

and, under this limitation

$$\theta = \alpha \cos\left(2\pi \frac{t}{\tau} + \delta\right) \quad (2)$$

α and δ being arbitrary constants, and τ , the periodic time, being given by

$$\tau = 2\pi(l/g)^{1/2} \quad (3)$$

Equations (2) and (3) provide the general solution of Eq. (1). On examination, they will be seen to represent an oscillation in angle, between limits $-\alpha \leq \theta \leq \alpha$, of periodic time τ which is independent of α . The constant δ specifies the displacement at the instant arbitrarily chosen as the zero of time; it is referred to as the "phase angle" of the motion related to this choice of time zero. The constant α is referred to as the (angular) amplitude of the oscillation. The fact that the periodic time (or frequency), in pendulum motion, is independent of amplitude (for small amplitudes) was first discovered empirically by Galileo, in 1581. A similar result is true for any motion (in angle, or any other coordinate) in which the acceleration towards the equilibrium position is proportional to the displacement, as required by Eq. (1). Motion having this character is "simple harmonic motion."

Pendulum motion has been considered in some detail because, historically, it was the first example of simple harmonic motion to be understood. For present purposes, however, more direct relevance might be seen in the motion of a load suspended by a helical spring. If such a load is displaced vertically, it executes simple harmonic motion about its position of EQUILIBRIUM. The extra upwards force acting on the load (by Hooke's law) being proportional to the extra extension of the spring, the acceleration of the load towards equilibrium is proportional to the displacement, as is required. But it is not easy, in practice, to approximate to the ideal of a massless spring supporting a massive load, and a theoretical account of the oscillations of a loaded spring which does not take count of the mass of the spring is oversimplified.

Simple harmonic motion is referred to as "isochronous" because, in a given case, the periodic time is the same whatever the amplitude. If the vibratory motion of an extended system were not similarly isochronous, there would be little point in claiming, as has already been done, that such motion, at its simplest, may be characterised by a periodic time which is unique. But the fact is that we have ample experience of real physical systems whose vibrations are of this character: it is natural to conclude, therefore, that the motion of each particle or element of a system so vibrating is simple harmonic motion. That is the reason why simple harmonic motion

has been considered in its own right, as a preliminary to the discussion of vibratory motion.

In vibratory motion, the restoring forces responsible for the individual motions of the constituent elements of a vibrating system may arise from tensions externally applied to the system (as with stretched strings and membranes), from elastic stresses developed internally as the system is deformed (as with rods vibrating longitudinally, transversely or torsionally; with air columns in organ pipes; with thin plates of regular shape), or occasionally in other ways (surface tension forces, for example are involved in the pulsation vibrations of thin films of liquid).

The simplest system of the first type is the stretched string. Imagine a uniform string, of length l and total mass ml , stretched under tension T between rigid supports. Let rectangular axes OX , OY be taken, along and at right angles to the length of the string in its undisturbed state, and let us consider possible transverse vibrations of the string in the XY plane. Suppose that all displacements (y) are very small compared with l , so that any over-all increase in the length of the string is negligible, and the tension may be regarded as constant throughout. Under these conditions, imagine that the string is in steady vibration, and at an arbitrary instant, let the transverse displacement vary along the length of the string as represented by the expression $y_0 = f(x)$. The instantaneous value of the curvature of the "displacement profile" of the string, at x , being given by $f''(x)$, the instantaneous magnitude of the force per unit length of the string about x is $Tf''(x)$. Under these imposed conditions, this force is essentially at right angles to OX and in the direction of y increasing—and, as a result, the acceleration of the element of the string around x is $(T/m)f''(x)$, instantaneously. Only if this acceleration is proportional to the instantaneous displacement at x , and if the proportionality constant is independent of x , will each element of the string execute simple harmonic motion of the same period—and only then will the postulated vibration of the string as a whole be well-defined. For such sustained motion, therefore, the necessary condition is

$$(T/m)f''(x) = -\mu f(x) \quad (4)$$

where μ is a constant, independent of x . When μ has been evaluated, ν , the frequency of the vibration, will likewise be known, for, in this case [compare Eqs. (1) and (3)]

$$\nu = \mu^{1/2}/2\pi \quad (5)$$

Now, the general solution of Eq. (4) is

$$f(x) = A \sin\left\{\left(\frac{\mu m}{T}\right)^{1/2} x + \epsilon\right\}$$

but, because the string is fixed at its two ends, so that $f(0) = f(l) = 0$,

$$\epsilon = 0$$

and

$$\left(\frac{\mu m}{T}\right)^{\frac{1}{2}} l = n\pi$$

n being a positive integer. For ν , therefore, we have [see Eq. (5)]

$$\frac{n}{2l} \left(\frac{T}{m}\right)^{\frac{1}{2}}$$

and, for $f(x)$,

$$f(x) = A \sin \frac{n\pi x}{l}$$

Again, since each element of the string executes simple harmonic motion of frequency ν , we have the following general expression for the displacement (y) of any element of string, given as a function of x , and the time t , namely

$$A \sin \frac{n\pi x}{l} \cos \frac{n\pi}{l} \left(\frac{T}{m}\right)^{\frac{1}{2}} t \quad (6)$$

Taking account of all possible values of n , Eq. (6) represents the whole series of simple modes in which the stretched string may vibrate. Generally, these modes are referred to as the "normal modes" of the system. Because the frequencies of these normal mode vibrations are the successive integral multiples of a basic frequency, $(1/2l)(T/m)^{\frac{1}{2}}$, they are said to form a "harmonic series," and the vibrations themselves are often described as the "first harmonic," "second harmonic," and so on. Alternatively, the first harmonic may be referred to as the "fundamental mode," the second harmonic as the "first overtone," and similarly throughout the series. It will be seen, from Eq. (6), that in the fundamental mode, the only points in the string which remain permanently at rest are the ends of the string; in the n th normal mode, in general, there are $(n-1)$ other points of permanent rest ("nodes") equally spaced between the two ends. Mid-way between the nodes are the points of greatest amplitude of motion ("anti-nodes").

By restricting consideration to the case in which the shape of the displacement profile of the string is the same at all instants throughout the vibration (being represented by $y_0 = f(x)$, above), we made certain that only normal mode vibrations should emerge from our analysis. But more complex sustained motion is possible in which the shape of the displacement profile continuously changes. To sustained motion of this general type, no single periodic time may be assigned. On the other hand it may be shown that every such motion may be regarded as resulting from the superposition of normal mode vibrations, the amplitudes and phases of the various modes being uniquely determinable in terms of the details of the actual motion as observed over a sufficient interval of time. From this point of view, the potentialities of the system

for sustained motion are completely given in Eq. (6) describing the normal modes.

With essentially linear systems, such as strings and rods and organ pipes, when the conditions at the two ends of the system are the same (as they are with the stretched string), and there is no internal constraint, the normal modes in general constitute a harmonic series; when the end conditions are different (as in a pipe which is closed at one end and open at the other) in general the even-numbered harmonics are not represented. In all cases, however, the theorem concerning the decomposition of the most complex sustained motion in terms of the normal modes applies.

As an example of elastic vibrations, we consider, briefly, the transverse vibrations of a thin, uniform rod, which is clamped at one end. For a displacement profile of arbitrary shape, the elastic forces brought into play across any section of the rod are proportional to the displacement of the rod at that section, and the unbalanced restoring force acting on a small finite element of the rod is similarly proportional to its displacement. But the proportionality constant relating restoring force per unit length to displacement will, in general, vary along the rod. Only for displacement profiles of certain shapes will this constant be the same over-all. These are the shapes of displacement profile corresponding to the normal mode vibrations, and it is the object of theory to identify them and to deduce the corresponding frequencies. Even an approximate theory is complicated; all that can be done here is merely to quote the ratios of the frequencies of the first three normal mode vibrations, namely 1:6.27:17.55. Very clearly, the normal modes in this case do not constitute a harmonic series: in musical parlance, the first overtone is some $2\frac{1}{2}$ octaves higher than the fundamental.

Throughout the above discussions, actual physical situations have been consistently idealized, in that it has been assumed, implicitly, that all periodic motion is sustained indefinitely. In fact, all such motion, unless energy is continually supplied from outside, gradually loses amplitude and dies away. This process is referred to as "damping," and the dissipation of energy which is its essential feature is generally, in mechanical systems, attributable to friction. Detailed consideration of damping is, however, beyond the scope of this article.

N. FEATHER

References

- Feather, N., "An Introduction to the Physics of Vibrations and Waves," Edinburgh, Edinburgh University Press, 1961.
- Lindsay, R. B., "Mechanical Radiation, New York, McGraw-Hill Book Co., Inc., 1960.
- Rayleigh (Lord), "Theory of Sound," Second edition, London, The Macmillan & Co., Ltd., 1894.

Cross-references: DYNAMICS, MUSIC, WAVE MOTION.

VISCOELASTICITY

Viscoelasticity is a material property possessed by solids and liquids which, when deformed, exhibit both viscous and elastic behavior through simultaneous dissipation and storage of mechanical energy. The *material constants* linking stress and strain in the theory of elasticity become *time-dependent material functions* in the constitutive equations of viscoelastic theory. At sufficiently small (theoretically infinitesimal) strains the behavior of viscoelastic materials is well described by the *linear theory of viscoelasticity* epitomized by the celebrated *Boltzmann superposition principle*. Expressed in its simplest form

$$\sigma(t) = \int_0^t Q(t-u)\epsilon(u) du$$

$$\epsilon(t) = \int_0^t U(t-u)\sigma(u) du$$

it states that the stress (or strain) at time t under an arbitrary strain (or stress) history is a linear superposition of all strains (or stresses) applied at previous times u multiplied by the values of a weighting function $Q(t)$ [or $U(t)$] corresponding to the time intervals $t-u$ which have elapsed since imposition of the respective strains (or stresses).

Considerable simplification of the viscoelastic relations is achieved by mapping them from the real t axis into the complex s plane through the Laplace transformation. The resulting transforms in s may be manipulated algebraically and then inverted to regain the time-dependent form. Thus the above convolution integrals become

$$\sigma(s) = Q(s)\epsilon(s) \quad \epsilon(s) = U(s)\sigma(s)$$

and $Q(t)$ and $U(t)$ can be seen to be the material functions representing the response to a unit impulse of strain or stress. The unit step response functions, the material functions under constant strain ϵ_0 and constant stress σ_0 , are the *relaxation modulus* and *creep compliance*, obtained as

$$G(t) = \sigma(t)/\epsilon_0 = \mathcal{L}^{-1} Q(s)/s$$

$$J(t) = \epsilon(t)/\sigma_0 = \mathcal{L}^{-1} U(s)/s$$

where \mathcal{L}^{-1} denotes inversion of the transform. For sinusoidal steady state strain $\epsilon(\omega)$ and stress $\sigma(\omega)$, there result the *complex modulus* and *complex compliance*

$$G^*(\omega) = \sigma(\omega)/\epsilon(\omega) = [Q(s)]_{s=i\omega}$$

$$J^*(\omega) = \epsilon(\omega)/\sigma(\omega) = [U(s)]_{s=i\omega}$$

whose real and imaginary parts are the *storage modulus* $G'(\omega)$, *storage compliance* $J'(\omega)$, *loss modulus* $G''(\omega)$, and *loss compliance* $J''(\omega)$. Their ratio is the *loss tangent*

$$\tan \delta(\omega) = G''(\omega)/G'(\omega) = J''(\omega)/J'(\omega)$$

Any of the above functions (or any other that may be derived, e.g., for constant rate of strain), if known over a sufficiently extended time scale

(or, equivalently, over a sufficient range of frequencies), provides complete information on the viscoelastic properties of a homogeneous isotropic material. As formulated above, the equations refer to deformation in simple shear. When dealing with combined stresses, as in viscoelastic stress analysis, the tensorial character of stress and strain must, of course, be taken into account. The equations for other types of deformation, e.g., dilatation or uniaxial extension, are analogous to those for shear. The relations between the frequency-dependent viscoelastic functions in shear, extension, and dilatation are the same as those between the elastic constants, while for the time-dependent functions these relations are valid between the s -multiplied Laplace transforms. Thus the relaxation modulus in extension $E(t)$ is related to $G(t)$ through the *time-dependent Poisson's ratio* $\nu(t)$ by

$$E(s) = 2G(s) [1 + \nu(s)]$$

and for $\nu(t) = 1/2$, $E(t) = 3G(t)$. Data obtained in shear and extension on incompressible bodies are thus readily combined. Moreover, if a viscoelastic function can be formulated analytically, it can be converted into any of the others, allowing combination of measurements made under different stress or strain histories. Use is made of this to extend the experimentally accessible time scale. Thus at short times, dynamic (frequency-dependent) measurements are more convenient than transient (time-dependent) measurements, and *vice versa* for long times.

Another important extension of the time scale is available through *time-temperature superposition*. An increase in temperature generally shortens the time necessary for the molecular rearrangement processes responsible for viscoelastic behavior. If all such processes are affected by temperature in the same way, the material is *thermorheologically simple*, and a change in temperature simply shifts the viscoelastic functions along the logarithmic time or frequency axis. For polymers (the typical viscoelastic materials) above the *glass transition temperature* (at which main chain motion effectively ceases), the shift factor a_T is given by an equation of the form (WLF equation)

$$\ln a_T = \frac{c_1(T - T_0)}{c_2 + T - T_0}$$

where T_0 is a suitably chosen reference temperature, and c_1 and c_2 are constants. Measurements made at different temperatures can thus be combined to yield a single *master curve*.

Interconversion of the viscoelastic functions is facilitated by the introduction of *spectral distribution functions*. The spectral functions may be derived conveniently by rewriting the Boltzmann superposition principle, given above as an integral operator equation, in the form of the *differential operator equation*:

$$[\Sigma p_m D^m] \sigma(t) = [\Sigma q_m D^m] \epsilon(t)$$

where $D' = d'/dr$. The Laplace transformation leads to

$$Q(s) = 1/U(s) = \sum q_m s^m / \sum p_m s^m$$

a definition of $Q(s)$ and $U(s)$ which is readily linked with the useful representation of viscoelastic behavior by *models* consisting of series-parallel combinations of springs (elastic or storage elements) and dashpots (viscous or dissipative elements). A parallel combination (Voigt model) is characterized by a *retardance* $U(s) = 1/(G + \eta s) = J/(1 + \lambda s)$, where $\lambda = \eta/G$ is the *retardation time*, η signifies viscosity, and $J = 1/G$. A series combination (Maxwell model) is characterized by a *relaxance* $Q(s) = 1/(J + 1/\eta s) = G\tau s/(1 + \tau s)$ where $\tau = \eta/G$ is the *relaxation time*.

Through the *combination rules*: "relaxances add in parallel, retardances add in series," $Q(s)$ and $U(s)$ are readily derived from a given more complex model as illustrated below.

These models are *conjugate*, i.e., with suitable choices for the parameters, they describe the same viscoelastic material. Representation of viscoelastic behavior by a (small) finite number of elements is often inadequate. Molecular theories of polymer behavior lead to models with an infinite number of parameters, characterized by a *discrete distribution* (or *line spectrum*) of *relaxation* or *retardation times*. These distributions are normally so closely spaced that they cannot be resolved experimentally. One can therefore define *continuous distribution functions*

$$Q(s) = \int_0^\infty G(\tau) \frac{\tau s}{1 + \tau s} d\tau = \int_0^\infty H(\tau) \frac{\tau s}{1 + \tau s} d\ln \tau$$

$$U(s) = J + \int_0^\infty J(\lambda) \frac{d\lambda}{1 + \lambda s} = 1/\eta s$$

$$J = \int_0^\infty L(\lambda) \frac{d\ln \lambda}{1 + \lambda s} = 1/\eta s$$

where $G(\tau)$ and $J(\lambda)$ are distributions of relaxation and retardation times respectively, and $H(\tau) = \tau G(\tau)$ and $L(\lambda) = \lambda J(\lambda)$ are their counterparts on the more convenient logarithmic time scale. The spectral functions may be derived from experimentally determined curves of the viscoelastic functions through any of several approximation methods.

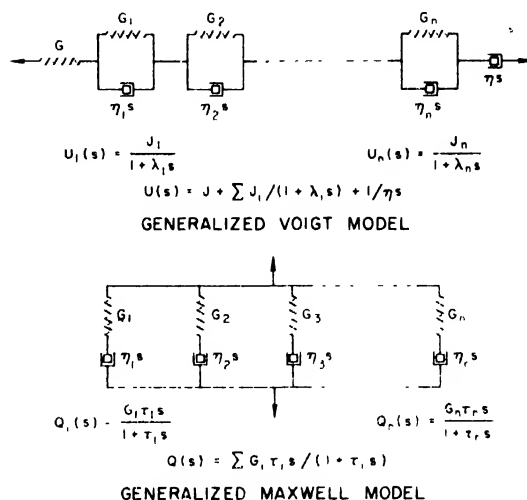


FIG. 1.

The various viscoelastic functions are illustrated below on the example of an uncross-linked amorphous polyvinyl acetate with a molecular weight of about 300 000. Data obtained in extension and shear at different temperatures were combined and interconverted. The functions are grouped to display their qualitative symmetries.

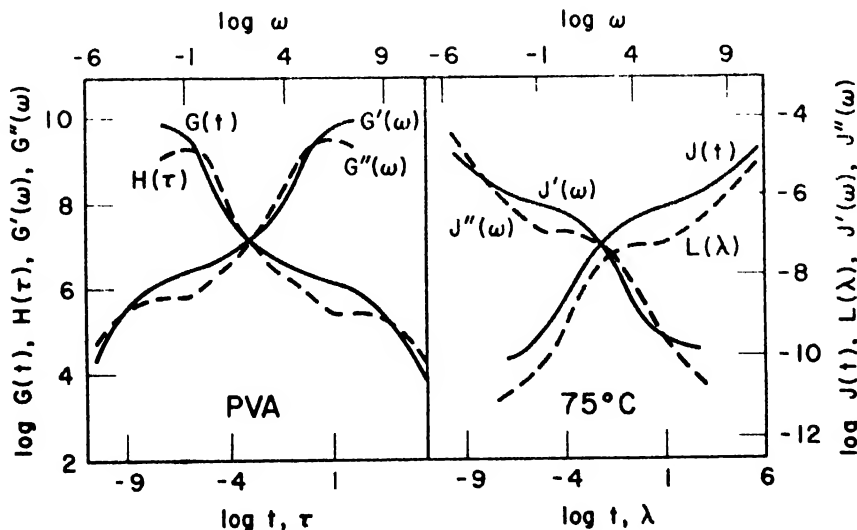


FIG. 2.

Singling out the relaxation modulus for a broad interpretation of the viscoelastic behavior of an uncross-linked polymer, one sees that at short times (in the *glassy region*), the modulus approaches an asymptotic value, the *glass modulus*. It then drops through the *glass-to-rubber transition region* and levels out somewhat in the *plateau region* reflecting the effect of molecular entanglements. At still longer times, it decays to zero through the *terminal region*. For cross-linked networks this region is absent (if there is no chemical degradation). In this case, the flow term $1/\eta s$ is missing from $U(s)$ above; an additive constant, the *equilibrium modulus* G_e , characterizing the (level) *rubbery region*, appears in $Q(s)$.

Current efforts in the field of viscoelasticity are directed chiefly towards the development of nonlinear (large deformation) theory, theories of anisotropic and inhomogeneous (semicrystalline and filled) systems, the solution of viscoelastic boundary problems (viscoelastic stress analysis), and the explanation of viscoelastic behavior on the molecular level (molecular theories).

N. W. TSCHOEGL

References

- Alfrey, T., and Gurnee, E. F., "Dynamics of Viscoelastic Behavior," in Eirich, F. R., Ed., "Rheology," Vol. 1, New York, Academic Press, 1956.
- Bueche, F., "Physical Properties of Polymers," New York, Interscience Publishers, 1962.
- Ferry, J. D., "Viscoelastic Properties of Polymers," New York, John Wiley & Sons, 1961.
- Leaderman, H., "Viscoelasticity Phenomena in Amorphous High Polymeric Systems," in Eirich, F. R., Ed., "Rheology," Vol. 2, New York, Academic Press, 1958.
- Nielsen, L. E., "Mechanical Properties of Polymers," New York, Reinhold Publishing Corp., 1962.
- Staverman, A. J., and Schwarzl, F., "Linear Deformation Behavior of High Polymers," in Stuart, H. A., Ed., "Die Physik der Hochpolymeren," Vol. 4, Berlin, Springer, 1956.
- Tobolsky, A. V., "Properties and Structure of Polymers," New York, John Wiley & Sons 1960.

Cross-references: ELASTICITY, POLYMER PHYSICS, RHEOLOGY, VISCOSITY.

VISCOSITY

Materials in general show two broad kinds of behavior with regard to their reaction when subject to an applied force. In one type, they deform until a position of equilibrium is reached, when no further change of shape takes place. Many solids show this kind of behavior. Alternatively, there is no permanent resistance to change of shape, and continuous deformation takes place for as long as the force is applied. The material is said to flow. Substances possessing this property are called fluids and may generally be thought of as either liquids or gases. The class also includes some materials, such as glass and pitch, which at

room temperature exhibit many of the characteristics normally associated with solids. When flow takes place in a fluid, it is opposed by internal friction arising from the cohesion of the molecules. This internal friction is the property of the fluid known as *viscosity*. It is clearly of great importance in many contexts; we may think, for example, of the flow of blood in the body, the flow of oil in pipe lines, the working of molten glass, and the process of lubrication. It may be used in the discrimination between streamlined and turbulent flow. It is also a useful property in providing information about the structure of complex organic molecules.

Definitions and Units. The formal definition of viscosity arises from the concept put forward by Newton that under conditions of parallel flow, the shearing stress is proportional to the velocity gradient. If the force acting on each of two planes of area A parallel to each other, moving parallel to each other with a relative velocity V , and separated by a perpendicular distance X , be denoted by F , the shearing stress is $\frac{F}{A}$ and the velocity gradient, which will be linear for a true liquid, is $\frac{V}{X}$.

Thus

$$\frac{F}{A} = \eta \cdot \frac{V}{X}$$

The constant η is known as the viscosity coefficient, dynamic viscosity, or viscosity of the liquid. The unit, expressed in dyne second per square centimeter, is known as the "poise," after Poiseuille who worked on viscosity in the early part of the nineteenth century. For practical purposes it is often more convenient to use a smaller unit, known as the centipoise, equal to a hundredth of a poise. It is useful to note that water at room temperature has a viscosity of approximately 1 centipoise. In many applications of viscometry, it is useful to note another quantity, called the kinematic viscosity, which is obtained by dividing the dynamic viscosity by the density. It is frequently denoted by the Greek letter ν , and the unit in which it is expressed is the "stokes," after Sir George Stokes, another pioneer worker in viscometry. In practical viscometry, viscosity is sometimes expressed in seconds; this is a measure of the time for a prescribed quantity of liquid to flow through a tube or aperture of defined dimensions.

In considering many problems of fluid motion, an important non-dimensional quantity is vd/ν , where v is the velocity of the fluid, ν its kinematic viscosity, and d a linear dimension, such as the diameter of a tube. This quantity is known as the Reynolds' number; when it is less than about 1000, flow is generally streamlined, whereas at values over 1000 it is turbulent.

Values for the viscosity of a number of common substances are given in Table 1.

Methods of Measurement. In measuring viscosity in the absolute sense, it is necessary to be

TABLE I. COEFFICIENT OF VISCOSITY IN POISES
(APPROXIMATE VALUES)

Water at 20°C	0.010
Mercury at 20°C	0.015
Carbon tetrachloride at 20°C	0.009
Olive oil at 20°C	0.84
Glycerine at 20°C	8.3
Golden Syrup at 20°C	1000
Glass at melting temperature	10 ³
Glass at working temperature	10 ⁷
Glass at annealing temperature	10 ¹³
Air at 0°C	0.00017
Carbon dioxide at 0°C	0.00014
Steam at 100°C	0.00013
Pitch at 15°C	10 ¹⁰

able to use the basic relationship between shearing stress and velocity gradient in order to derive a relationship between the measurable quantities involved in flow in an apparatus of a particular shape. For example, it can be shown that the volume Q of a liquid of viscosity η flowing per unit time through a tube of length l and internal radius a , when the pressure difference between the ends is P , is given by

$$Q = \frac{\pi Pa^4}{8l\eta}$$

Capillary viscometers enable all the variables in this equation to be measured and the viscosity to be determined.

Similarly, a viscometer can be constructed in which one cylinder rotates inside another with the liquid under test in the annular space between the two. Measurement of the angular velocity and the applied torque, together with a knowledge of the dimensions of the apparatus, enables the viscosity to be calculated.

Another method involves the measurement of the rate of fall of a sphere in a column of liquid. If the liquid has viscosity η and density ρ , and the sphere has radius r and density σ , then the velocity of fall is

$$v = \frac{2gr^2(\sigma - \rho)}{9\eta}$$

a relationship commonly known as Stokes's law.

There are also a number of empirical methods that can be used to obtain comparative values, the instruments being calibrated by using liquids of known viscosity.

Variation with Temperature. The viscosity of liquids decreases rapidly with increasing temperature. For example, the viscosity of molten glass may be halved by raising the temperature 30°C. On the other hand, the viscosity of gases increases with temperature. This feature is of practical importance in two senses. In all methods of measurement strict temperature control is necessary and when results are quoted, it is essential to state the temperature to which they refer.

The variation of viscosity with temperature is of great significance in the problem of lubrication, where oils may have to operate over widely

different conditions. So-called viscostatic oils have a low temperature coefficient and can thus be used over a wide temperature range. As an indication of this property, the oil industry uses an empirical number known as the "viscosity index"; the higher the viscosity index, the less is the variation of viscosity with temperature.

Molecular Weight Determinations. The viscosity of organic materials in solution was suggested by Staudinger as a useful index of their molecular structure, since the flow properties would be influenced by the size and shape of the molecules. For example, some high polymers give appreciable increases of viscosity even at low concentrations on account of the effect of randomly coiled long-chain molecules. Generally speaking, the higher the molecular weight, the greater is the increase in viscosity for a given weight in the solution. The values of molecular weight obtained by this method are not absolute, but they depend on the establishment of empirical relations by measurements on substances of known molecular weight in a given series (see MOLECULAR WEIGHT).

Non-Newtonian Systems. A large number of industrially important materials do not obey the simple Newtonian relationship between shearing stress and shearing rate. In some cases the viscosity varies with the shearing rate, and in others, it varies with time when the shearing rate is constant. An important group of such materials are known as thixotropic substances—they become thinner when stirred. Paint, suspensions of clay in water, and many other substances behave in this way, and the property has considerable industrial significance. In the case of paint, for example, the low viscosity when brushed in a thin film makes for easy application, whereas the high viscosity under low shearing stresses enables the film to be retained on vertical surfaces. Some substances thicken up on stirring, and there are a number of other classes of behavior (e.g. VISCOELASTICITY) that constitute departure from Newtonian laws. The study of such anomalous systems is an important branch of modern RHEOLOGY.

A. DINSDALE

References

- Newman, F. H., and Searle, V. H. L., "The General Properties of Matter," London, Ed Arnold, 1962.
- Dinsdale, A., and Moore, F., "Viscosity and Its Measurement," London, Chapman and Hall, and New York, Reinhold Publishing Corp., 1963.
- Perry, J. H., "Chemical Engineers Handbook," 4th Ed., New York, McGraw-Hill Book Co., 1963.
- Flory, P. J., "Principles of Polymer Chemistry," New York, Cornell University Press, 1963.
- Scott-Blair, G. W., "A Survey of General and Applied Rheology," London, Pitman, 1949.
- Eirich, F. R., Ed., "Rheology, Theory and Applications," Three volumes, New York, Academic Press, 1958.

Cross-references: FLUID DYNAMICS, MOLECULAR WEIGHT, RHEOLOGY, VISCOELASTICITY.

VISION AND THE EYE

The meaningful experience of vision requires light, the eye, and a conscious observer (animal or human) having an intact *visual system*. In its gross aspects, the visual system includes the eyes, the *extraocular muscles* which control eye position in the *bony orbit* (eye socket), the optic and other nerves that connect the eyes to the brain, and those several areas of the brain that are in neural communication with the eyes. This summary will stress the informational aspects of human vision; it should be realized however that no visual system could function without its *protective mechanisms* (tears and eyelids, especially) or if its normal metabolism (mediated through the vascular supply of eye and brain) were seriously interfered with.

The visual system is particularly well adapted for the rapid and precise extraction of spatial information from a more-or-less remote external world; it does this by analyzing, in ways that are as yet imperfectly understood, the continuously changing patterns of radiant flux impinging upon the surfaces of the eyes. Much of this light is reflected from objects which must be discriminated, recognized, attended to, and/or avoided in the environment; this ability transcends enormous variations in intensity, quality, and geometry of illumination as well as vantage point of the observer. A block diagram of the visual system is given in Fig. 1.

Although image formation in the eye is importantly involved, the analogy between eye and photographic camera has been badly overworked and tends to create the erroneous impression that little else is needed to explain how we see. Image formation is greatly complicated by the movement of the eyes within the head, and of both eyes and head relative to the external sea of radiant energy. Such visual input is ordinarily sampled by discrete momentary pauses of the eyes called *fixations*, interrupted by very rapid ballistic motions known as *saccades* which bring the eyes from one fixation position to the next. Smooth movements of the eyes can occur when an object having a predictable motion is available to be followed. A large body of evidence suggests that the visual input is processed by the brain in "time frames" of about 100 msec, although the

peripheral parts of the visual system operate with much shorter time constants than this.

Each eye controls many important functions within one mobile housing: it is a device to form an image upon a vast array of light sensitive *photoreceptors*, but it also contains systems to dissect, encode, and transmit information derived therefrom. A crosssection of the human eye is shown in Fig. 2. The primary refracting surface is the *cornea*, a complex yet transparent structure which admits light through the anterior part of the outer surface of the eye. The *iris* contains muscles which alter the size of the entrance port of the eye, the *pupil*. The *crystalline lens* has a variable shape, under the indirect control of the *ciliary* muscle. Since it has a refractive index higher than the surrounding media, it gives the eye a variable focal length, allowing *accommodation* to objects at varying distances from the eye. The iris muscles and the ciliary muscle, known collectively as the *intraocular* muscles of the eye, are controlled by impulses having their origins in separate but interacting centers in the brain stem. These brain centers also receive nerve impulses from the eye. These loops, and those involving the extraocular musculature and thus eye position, have some of the properties of nonlinear servosystems, and have been actively investigated as such.

Much of the remainder of the eye is filled with fluids and materials under pressure, which help the eye maintain its shape. The *aqueous humor*—thin, watery, and continuously being replaced—fills the *anterior chamber* between cornea and lens. The *vitreous humor*—thinly jellylike and of very low metabolism—fills the majority of the eye's volume. The image produced through these structures is formed upon the *retina* at the back of the eye. The retinal image is very small, because the eye itself is small and has a short posterior focal length of about 19 to 23 mm, depending upon accommodative state. The retinal image has a point-spread function on the order of two to three minutes of arc, corresponding to about 10μ on the retina for ideal conditions. These conditions include a 2 to 3 mm pupil, monochromatic light, optimal accommodation and a normal, young, and healthy eye. This quality approaches, but is somewhat worse, than that produced by diffraction-limited imagery

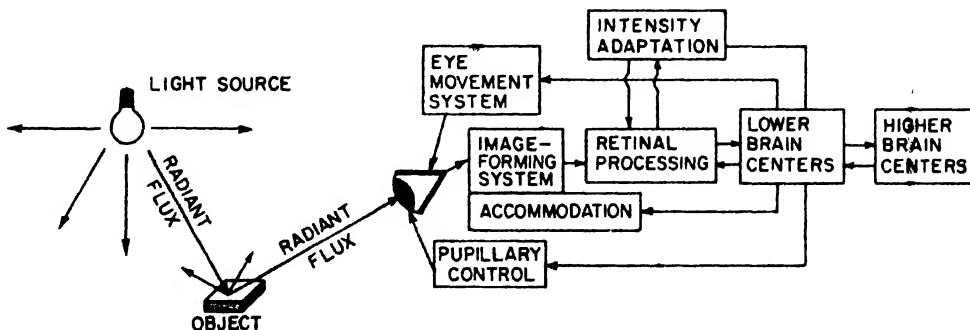


FIG. 1. Block diagram of the functional components of the visual system.

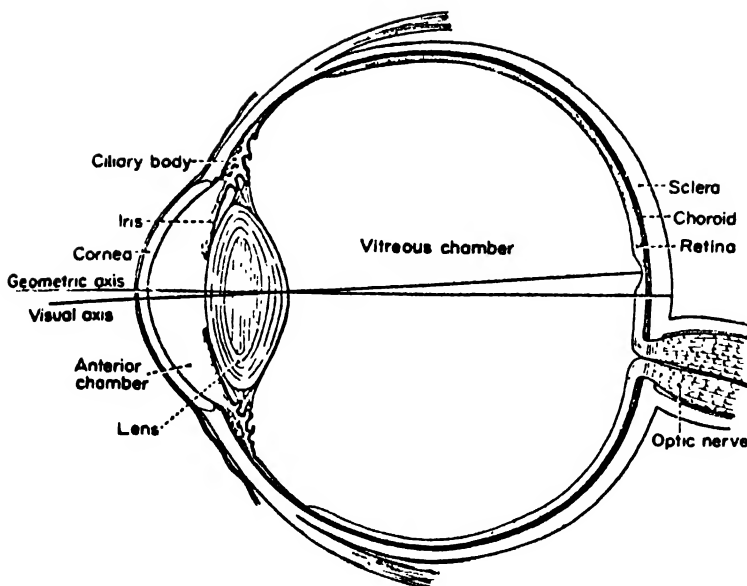


FIG. 2. Horizontal cross section of the right eye of the human.

in an ideal optical system. The retinal image is always in motion: even during the best efforts at steady fixation, there exists an irreducible tremor of the eye whose high-frequency components are in the 20 to 30 second-of-arc range, with larger drifting and saccadic movements up to 5 minutes of arc. It is possible to eliminate this residual motion by various optical techniques; such stabilization usually results in a total loss of vision, providing an elegant demonstration that the visual system responds primarily to *changes* in light patterns, rather than to steady states. Electrophysiological evidence from animals amply confirms this.

The *retina* is a thin structure of extreme complexity. It is considered embryologically to be a displaced part of the brain, and it is of clinical importance as the only part of the central nervous system that can be directly observed in the intact living subject. The receptors, the *rods* and *cones*, line the back surface of the retina, in immediate contact with a dark layer (the choroid) which helps to nourish the receptors and to prevent multiple reflection of light. There are about 125 000 000 receptors in each human eye, of which only about 5 per cent are cones. The cones are however of an importance disproportionate to their relative number: in particular there is a small central bouquet of about 2000 of them, located in a rod-free depression of the retina known as the *fovea centralis* where they are packed together into an hexagonal array having a density of about 150 000/mm²; these are capable of dissecting the finest details of the optimal retinal image. This process is aided by the lateral displacement of other retinal structures through which light must pass to reach the cone receptors. Moreover, this is the area of the retina where images have the highest

attention value and which "projects" to a disproportionately huge area of the visual brain; the extraocular muscles move the eye more or less automatically, in the act of fixation, to put objects of interest into this region where their details can be most critically appreciated, while the accommodative mechanism alters the shape of the lens to produce the sharpest possible image in this region. The cones, including those in the fovea, function only at high luminance levels (approximately, above .01 candela/m²), below which they are functionally blind and the rods take over. Thus the retina contains two systems intermixed: (a) the cone system (photopic), good for high-acuity vision, which also mediates all color vision; (b) the achromatic rod system, which has relatively poor spatial resolving power, but very high sensitivity.

The rods and cones are synaptically connected to the *bipolar cells*, which in turn relate to the *ganglion cells*, whose axons constitute the optic nerve fibers. There are also rich horizontal connections among the receptors, among the bipolar cells, and among the ganglion cells. In addition, there is a high degree of convergence: the 125 000 000 receptors ultimately feed into only 1 000 000 nerve fibers of the flexible optic nerve, which therefore constitutes the principal "bottleneck" of information flow in the visual system. The convergence ratio for the fovea is about 1:1, helping to preserve the high-detail vision of this region, while in the peripheral retina this ratio is many thousands to one, leading to high sensitivity at the sacrifice of resolving power.

The pathways from retina to brain are by no means independent, including those emerging from the central fovea. The horizontal interconnections are utilized to allow inhibitory processes to sharpen the "neural image" by a

process of border enhancement, but much more complicated preprocessing of information occurs also. It is abundantly clear that the brain does not receive a replica of what is on the retina, although a spatial isomorphism between retina and brain does exist; rather, the messages sent to the brain tend to carry information that is already processed in complex ways to make efficient use of the limited communications pathways between eye and brain in an adaptively significant manner.

Because the two eyes are located in slightly different places in space, a disparity of the two retinal images results. Rather than to produce a blurred or confused picture, this *retinal disparity* results in the appearance of *stereoscopic depth*. Such depth judgments are remarkably precise, consistent with the findings that all but the smallest eye movements and accommodative adjustments are highly correlated between the two eyes, and that neural units in the visual brain are precisely connected, by way of intermediate synapses, to optically corresponding areas.

The normal eye exhibits a large amount of chromatic aberration which is not normally noticeable. There are at least two reasons for this: (a) the cone receptors exhibit a directional sensitivity which reduces the visual effectiveness of light entering the marginal zones of the pupil; (b) the visual system has a remarkable capacity to adapt to systematic distortions of almost any kind which do not carry useful information from the external world. For example, observers learn with practice to compensate for the effects of gross visual displacement caused by prisms placed before the eyes, and are no longer able to see the chromatic fringes produced by such prisms. Removal of the prisms produces reappearance of chromatic fringes and an apparent displacement in the opposite direction. The explanation of such effects is not simple: in this example, the adaptation to displacement is probably kinesthetic rather than visual, but the adaptation to fringes is almost certainly confined to the visual system. Related to this are many entoptic phenomena that are seldom perceived: (a) the shadows of the retinal blood vessels, which are in front of the receptors; (b) the blind spot in the visual field, caused by the receptor-free optic disc, large enough to contain 200 images of the moon; (c) "floaters," usually shadows of debris in the vitreous humor, clearly visible if attended to against bright, uniform surfaces such as the sky; (d) fleeting specks of light probably caused by the movements of corpuscles within the retinal blood vessels; (e) Maxwell's spot, probably corresponding to the region of the macular pigment, and many others.

The initial nonoptical event in the visual process is the absorption of single light quanta by single molecules of visual photopigment, of which millions are located in each rod or cone. Under ideal conditions, as few as a half-dozen of these elemental events within fairly broad bounds of time and area, are sufficient to lead to a visual sensation. The visual photopigment contained in

the rods is *rhodopsin*, having a peak sensitivity at about 505 nm. It has been much studied and is found in most animals including man. Absorption of light by rhodopsin probably produces graded potentials at the receptors that trigger all-or-none nerve impulses by the time the ganglion cells are activated, if not before. The exact mechanisms whereby light absorption gives rise to receptor potentials (and these to nerve impulses), although under active investigation, cannot be said yet to be satisfactorily understood.

Color vision depends upon the existence of three classes of visual photopigment, all different from rhodopsin, housed in different proportions in different classes of cone receptors. When two fields of light that are physically different look exactly alike in color (*metameric matches*), it is probable that the rate at which light is being absorbed in the three classes of cone photopigment is the same from both fields, although this is not yet definitely established. The perception of color clearly involves the higher levels of the visual system as well.

Another important property of the eye is its adaptation to intensity by means of which the eye changes its gain and other characteristics, enabling it to respond discriminatively over a stimulus intensity range of about ten billion to one. At one time, it was felt that the bleaching of photopigments was primarily involved in this process; recent evidence indicates that this plays only a minor role and the true mechanisms are numerous and include changes in the organizational properties of retinal networks.

ROBERT M. BOYNTON

References

- Brindley, G. S., "Physiology of the Retina and the Visual Pathway," Baltimore, Williams and Wilkins Co., 1960
- Davson, H., Ed., "The Eye," New York, Academic Press, 1962, 4 vols
- Helmholtz, H. von., "Physiological Optics," translated by J. P. C. Southall, Optical Society of America, 1924, New York, Dover Publications, Inc., 1962
- LeGrand, Y., "Light, Colour, and Vision," translated by R. W. G. Hunt, J. W. T. Walsh and F. R. W. Hunt, New York, John Wiley & Sons, 1957.
- Piéron, H., "The Sensations," translated by M. H. Pirenne and B. C. Abbott, London, F. Müller, 1952.
- Pirenne, M. H., "Vision and the Eye," London, Chapman & Hall, 1948.
- Polyak, M., "The Vertebrate Visual System," Chicago, University of Chicago Press, 1957.
- Smelser, G. K., Ed., "The Structure of the Eye," New York, Academic Press, 1961.
- Weale, R. A., "The Eye and Its Function," London, Hattori Press Ltd., 1960.
- Wolff, E., "The Anatomy of the Eye and Orbit," Philadelphia, Blakiston Co., 1948.

Cross-references: COLOR; LENS; LIGHT; LUMINANCE; OPTICS; GEOMETRICAL; OPTICS, PHYSICAL; PHOTON; REFLECTION; REFRACTION.

VITREOUS STATE

The vitreous state, or the glassy state, is a special metastable condition in which a substance can exist. A material in the vitreous state may be termed a noncrystalline solid or a rigid liquid. The attainment of the vitreous state is illustrated by the volume-temperature relationship of an ideal system in Fig. 1. When a liquid is cooled slowly, crystallization will usually occur when the temperature reaches the melting point T_m as described by the path a-b. In the case of a complex liquid, crystallization will similarly commence when the temperature reaches the liquidus. Crystallization is generally considered to take place by a two-step mechanism, namely, nucleation and then crystal growth. In the absence of crystallization, a liquid can be supercooled below T_m , and the cooling path is now described by a-c. Since the fluidity of a liquid generally exhibits an exponential dependence on temperature, progressive undercooling along a-c will be accompanied by a rapid increase of viscosity. If crystallization has still not occurred when point d is reached, the viscosity of the liquid will have reached 10^{14} poises. (By comparison, the viscosity of glycerol at room temperature is only 10 poises.) For a particular cooling rate, say 5 deg/min, the cooling curve will now follow d-e. At any temperature below that corresponding to the point d, the material is said to be in the vitreous state. It now has the rigidity of the corresponding crystalline solid, but its structure is still devoid of long-range order similar to that of the parent liquid. A substance in the vitreous state is called a glass.

At the point d, the viscosity of the supercooled liquid is 10^{14} poises. At such high viscosity, molecular motions are retarded since the re-

laxation times are of the order of minutes and hours. The time taken for experimental measurements may now be actually less than the time needed to attain internal equilibrium. Thus for a particular cooling rate, an inflection will occur at d. For a slower cooling rate, d-e will be displaced to the first dashed line below it. For an even slower cooling rate, the inflection will occur at point f; the cooling curve will now be described by the lowest dashed curve. Similar behavior is shown by the heat content-temperature relationship of the system. The inflection points d and f, and hence the specific volume of a glass at the lower temperatures, are dependent on the cooling rate. The temperature region over which this inflection occurs is termed the glass-transition temperature T_g . Although the supercooled liquid at above T_g is metastable with respect to the crystalline phase, it is in internal thermodynamic equilibrium. A glass, however, is generally considered not to be in internal thermodynamic equilibrium.¹

Many substances are easily rendered into the vitreous state by supercooling the melt. These include organic liquids such as toluene and the alcohols, polymeric materials, fused salts of the system $\text{Ca}(\text{NO}_3)_2$ KNO_3 and many silicates, borates, phosphates and their mixtures.² In general the ease of glass formation via the liquid is dependent on the crystallization kinetic constants as well as the cooling rate. Although the crystallization kinetic constants of metallic systems, for instance, are not favorable for glass formation, the vitreous state is attainable by very rapid cooling. The condensation of vapor at low temperatures can also yield noncrystalline solid phases. Little is known about the relationship between such phases and the more common glassy phase obtained from the cooling of the

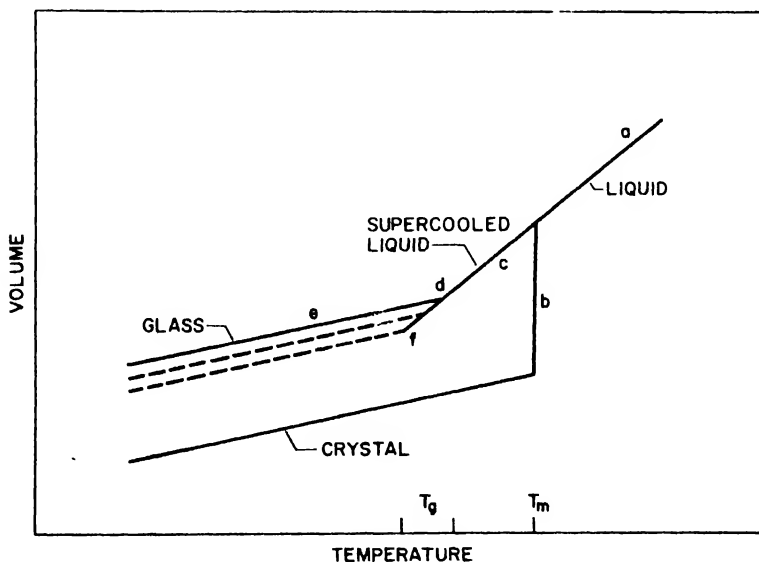


FIG. 1. Volume-temperature relationship of an ideal glass-forming system.

melt. Present theoretical interests are centered on the nature of the glass transition at T_g . Two important questions are: (a) Is the glass-transition region an iso-free volume state for all substances? (b) Is there a theoretical lower temperature limit to T_g ? No unambiguous answers to these questions are presently available.

JOHN D. MACKENZIE

References

- Kauzmann, W., "The Nature of the Glassy State and the Behavior of Liquids at Low Temperatures," *Chem. Rev.* **43**, 219 (1948).
2. Mackenzie, J. D., "Modern Aspects of the Vitreous State," Vol. 1, London, Butterworths, 1960.

Cross-references: CONDENSATION, CRYSTALLIZATION, CRYSTALLOGRAPHY.

W

WAVE MECHANICS

Introduction. It has been known since the early part of the present century that systems of atomic dimensions do not obey the laws of classical mechanics, as formulated by Newton. A new mechanics has had to be constructed, and this is called *quantum mechanics*.

There were initially two apparently different formulations of quantum mechanics, one called **MATRIX MECHANICS** and the other *wave mechanics*. It was ultimately shown that the two formulations are completely equivalent and complicated quantum mechanical problems are nowadays usually solved by a combination of both. However, there are still many applications of quantum mechanics which can be treated by purely wave mechanical methods, and as these generally make a more direct appeal to physical intuition, it is perhaps worth while to give an account of wave mechanics quite separate from matrix mechanics.*

Waves and Particles. The work of Planck, in 1900, on thermal radiation, and of Einstein, in 1905, on the photoelectric effect, suggested that light must, for some purposes, be regarded as consisting of particles, called *light quanta* or *photons*. The energy of a photon is given by

$$E = h\nu \quad (1)$$

and its momentum by

$$p = h/\lambda \quad (2)$$

where ν is the frequency and λ the wavelength of the light, and h is Planck's constant, which has the value 6.624×10^{-27} erg sec. On the other hand, diffraction phenomena, for example, can only be explained on the assumption that light consists of waves, so that neither the corpuscular nor the wave theory is completely satisfactory. In fact, it must simply be accepted that light can behave *either* as particles *or* as waves. The nature of a given experiment will emphasize one aspect or the other, and the relations between the two aspects are given by Eq. (1) and (2).

In 1924, de Broglie went further and suggested that *any* moving particle, with mass m and

speed v , will in some experiments display wave-like properties, the wavelength being given by Eq. (2), with $p = mv$. This purely theoretical suggestion received experimental confirmation in 1927, when Davisson and Germer observed the diffraction of a beam of electrons by a crystal of nickel, and similar results have since been obtained with beams of other kinds of particles, including atoms and molecules.

De Broglie's theory was the beginning of wave mechanics, but its further development was due to Schrödinger, who showed, in 1926, how the theory could be used to account for the existence of stationary states of atoms.

In 1911, Rutherford had proposed that an atom consists of a small nucleus surrounded by a planetary system of electrons. The electrons are continually accelerated; hence, according to classical electromagnetic theory, there should be a continual emission of radiation, accompanied by a diminution in the size of the electronic orbits and consequent increase in frequency. The existence of line spectra, that is, the emission of radiation in a discrete series of frequencies, and indeed the stability of matter itself, proves that this is not the case, and in 1913 Bohr suggested that an atom can exist in any one of a set of so-called *stationary states*, each with a definite energy, and with a finite energy difference between one and the next. Radiation is emitted only when an atom passes from a stationary state with energy E_1 to a state of lower energy E_2 , and the frequency ν of the radiation is given by

$$h\nu = E_1 - E_2 \quad (3)$$

Bohr attempted to graft this idea on to the classical picture of electrons describing planetary orbits about the nucleus by postulating that only certain orbits were permissible. This theory successfully explained the line spectrum of hydrogen and had several other partial successes, but it was soon superseded by the more revolutionary but also more versatile quantum mechanics.

The Schrödinger Equation. According to de Broglie's theory a freely moving particle should be represented by a wave of the form

$$\Psi(x, t) = A \exp \left[2\pi i \left(\frac{x}{\lambda} - \nu t \right) \right]$$

* According to a recent work of Dirac, the validity of wave mechanics does not extend to quantum field theory, but it remains perfectly adequate in the domain of atomic and molecular physics.

where A is a constant. Using Eq. (1) and (2), this becomes

$$\Psi(x, t) = A \exp[i(px - Et)/\hbar] \\ \psi(x) \exp(-iEt/\hbar) \quad (4)$$

where $\hbar = h/2\pi$ and

$$\psi(x) = A \exp(ipx/\hbar) \quad (5)$$

is the time-independent part of the wave (we shall deal with the time-dependent factor later).

Differentiation shows that

$$\frac{d^2\psi}{dx^2} = -\frac{p^2}{\hbar^2}\psi \quad (6)$$

However, the energy of the particle is

$$E = \frac{p^2}{2m} + V$$

the first term on the right-hand side being the kinetic energy and the second term, the potential energy, which is constant if the particle is moving under no forces. Equation (6) thus becomes

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}(E - V)\psi = 0 \quad (7)$$

Although this equation has been plausibly derived only for a free particle, Schrödinger assumed that it also applies to a particle moving under a force, so that the potential energy V is a function of x . It is known as the *Schrödinger equation* or the *wave equation* for a particle moving in one dimension, and the function $\psi(x)$ is called the *wave function*. We will consider the interpretation of ψ below, but meanwhile let us note that ψ is to be *single-valued, continuous, smooth* [except at infinities of $V(x)$] and *finite everywhere*. Also, although $\psi \neq 0$ is always a solution of Eq. (7), it is not permitted as a wave function.

Particle in a One-dimensional Box. That the Schrödinger equation does lead to discrete energy values, at least when classical mechanics would predict a limited range of motion or a *bound* particle, can easily be seen by considering a particle moving in one dimension between infinite potential barriers—this is known as a one-dimensional box.

Suppose that the potential energy is zero from $x = 0$ to $x = L$, and infinite everywhere else. In the region of zero potential energy, Eq. (7) becomes

$$\frac{d^2\psi}{dx^2} + \frac{2mE}{\hbar^2}\psi = 0 \quad (8)$$

with the general solution

$$\psi = A \cos \sqrt{\frac{2mE}{\hbar^2}} x + B \sin \sqrt{\frac{2mE}{\hbar^2}} x \quad (9)$$

The condition of finiteness on ψ (which we will not consider in detail) demands that it vanish in

the region of infinite potential energy, and because of the condition of continuity, we must then have $\psi(0) = \psi(L) = 0$. The only solutions in the region of zero potential energy satisfying these boundary conditions are of the form

$$\psi_n = B_n \sin \frac{n\pi x}{L}, \quad n = 1, 2, 3, \dots \quad (10)$$

corresponding to the energy values E_n given by

$$\sqrt{\frac{2mE_n}{\hbar^2}} = \frac{n\pi}{L} \\ \text{or} \quad E_n = \frac{n^2\pi^2\hbar^2}{2mL^2} \quad (11)$$

($n = 0$ gives $\psi \equiv 0$, which is not allowed).

The energy E can thus have any one of an infinite set of discrete values, corresponding to the integral values of n , but no other value. These values are called the *eigenvalues* of the Schrödinger equation or *energy levels* of the system, and the corresponding wave functions are called the *eigenfunctions* of the equation and represent the stationary states of the system.

Operators. The Hamiltonian $H(x, p)$ of the one-dimensional system we have been considering is simply the energy expressed in terms of the momentum p and coordinate x , and the *equation of energy* is

$$H(x, p) = \frac{p^2}{2m} + V(x) = E \quad (12)$$

If we formally convert this into an *operator* equation by letting

$$p \rightarrow \frac{\hbar}{i} \frac{d}{dx} \quad (13)$$

(the coordinate x becomes the operator "multiply by x ", but this is trivial), and allow both sides to operate on a function $\psi(x)$, we obtain

$$\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V\psi = E\psi \quad (14)$$

which, when rearranged, is seen to be the Schrödinger equation [Eq. (7)].

The representation of dynamical variables by operators is fundamental to quantum mechanics, but the choice of operators made above is not unique. It may be confirmed by differentiation that

$$\frac{\hbar}{i} \frac{d}{dx} (x\psi) = x \frac{\hbar}{i} \frac{d\psi}{dx} + \frac{\hbar}{i} \psi$$

and, using Eq. (13), this gives the operator equation

$$px - xp = \frac{\hbar}{i} \quad (15)$$

The operator on the left is called the *commutator* of p and x , and is generally written $[p, x]$. Equation (15) is the basic equation of matrix mechanics and existed before Schrödinger's

equation, p and x then being represented by MATRICES. This shows in an elementary way the close connection between wave mechanics and matrix mechanics.

Particles in Three Dimensions. Equation (13) gives the clue to the extension of Schrödinger's equation to systems of one or more particles moving in three dimensions. We first write down the Hamiltonian of the system and transform each Cartesian component of momentum into a differential operator like Eq. (13) with respect to its corresponding (canonically conjugate) coordinate. For example, for a single particle moving in three dimensions, the Hamiltonian is

$$H(x, y, z, p_x, p_y, p_z) = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2) + V(x, y, z)$$

where p_x, p_y, p_z are the Cartesian components of the momentum, and the equation of energy is

$$H(x, y, z, p_x, p_y, p_z) = E \quad (16)$$

If we transform the latter into an operator equation by writing

$$p_x = \frac{h}{i} \frac{\partial}{\partial x}, \quad p_y = \frac{h}{i} \frac{\partial}{\partial y}, \quad p_z = \frac{h}{i} \frac{\partial}{\partial z} \quad (17)$$

(partial derivatives are now required as several variables are present), and operate upon a function $\psi(x, y, z)$, we obtain the equation

$$-\frac{h^2}{2m} \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} \right) + V(x, y, z)\psi = E\psi \quad (18)$$

which is the Schrödinger equation of the system. The essential correctness of this equation has been amply demonstrated by many successful applications (this is, of course, the *only* justification of any of the postulates of quantum mechanics). In particular, in the case of the hydrogen atom, where there is a single electron moving under the Coulomb attraction of a single proton (which may as a good approximation be considered to be at rest), the equation can be solved analytically, and the energy levels so found, together with the Bohr frequency rule [Eq. (3)], correctly give the line spectrum of hydrogen.

The extension to many particles is straightforward; in the Hamiltonian of the system, the substitutions of Eq. (17) are made for the momentum components of each particle of the system, and the equation of energy is thus transformed into the Schrödinger equation. Again the equation has had many successful applications, an early one being the calculation of the energy of the normal helium atom by Hylleraas in 1930, which the earlier Bohr theory failed to do. Unfortunately, owing to the interaction of the electrons, the equation cannot be solved analytically for many-electron systems and resort must be made to approximate methods of solution which have become extremely complicated in recent years.

Interpretation of the Wave Function. Let us consider the Schrödinger equation for a single particle moving in three dimensions. This is a linear equation, so that if ψ is a solution, so also is $C\psi$, where C is any constant. This means that (in most cases) we can choose ψ so that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\psi(x, y, z)|^2 dx dy dz = 1 \quad (19)$$

The function ψ is generally complex and $|\psi|$ is its *modulus* ($|\psi|^2 = \psi^* \psi$, where ψ^* is the complex conjugate of ψ). If it satisfies Eq. (19), the function ψ is said to be *normalized*.

The interpretation of ψ demands that it be normalized. If this is not possible, owing to the divergence of the integral in Eq. (19), as is the case with a free particle, then normalization has to be affected artificially by enclosing the system in a large box.

It was proposed by Born, in 1926, that if the particle is in a stationary state, with normalized wave function ψ , then $|\psi|^2$ may be interpreted as a *probability density*, such that

$$|\psi(x, y, z)|^2 dx dy dz$$

is the *probability* of finding the particle in the small volume element $dx dy dz$ at the point (x, y, z) .

It is clear from this why normalization is necessary—Eq. (19) expresses the fact that the probability of finding the particle *somewhere* is unity.

An important extension of this interpretation, which we will not justify in detail, relates to the *average* or *expectation value* \bar{f} of a dynamical variable f (such as momentum) for a system in a state ψ , i.e., the average of a large number of measurements of f made on a system in this state. It is found that

$$\bar{f} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \psi^* f_{op} \psi dx dy dz \quad (20)$$

where f_{op} is the quantum mechanical operator representing f . Here we have restricted ourselves to a single-particle system, but the same applies to systems of any number of particles, so long as the integral is a multiple integral taken throughout the full range of all the variables appearing in ψ .

The interpretation of ψ , or rather of $|\psi|^2$ (ψ itself has no physical significance), which we have presented above, gives rise to a fundamental and important difference between classical mechanics and quantum mechanics. According to the latter, there is a finite probability of finding a particle in regions of space where its presence would be forbidden by classical mechanics—regions where its kinetic energy would be negative.

Let us take as an example the linear harmonic oscillator. Suppose a particle of mass m is moving along the x axis under the action of a force $-kx$ (k is positive) directed towards the origin. Classically, the particle executes simple harmonic motion with frequency $\omega/2\pi$, where

$\omega = \sqrt{k/m}$. It is easy to see that the potential energy is $V(x) = \frac{1}{2}m\omega^2 x^2$, if the zero of potential energy is taken to be at $x = 0$, so that the Schrödinger equation [Eq. (7)] becomes

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2} \left(E - \frac{1}{2}m\omega^2 x^2 \right) \psi = 0 \quad (21)$$

It may be verified by substitution that the function

$$\psi_0 = \exp(-m\omega x^2/2\hbar) \quad (22)$$

satisfies this equation, provided E has the value

$$E_0 = \frac{1}{2}\hbar\omega \quad (23)$$

This is, in fact, the *ground state* of the oscillator, that is to say, E_0 is the lowest energy level.

Now, according to classical mechanics, the particle, having energy E_0 , would be confined to a region of the x axis whose limits are given by setting $V(x) = E_0$, that is, $x = \pm \sqrt{\hbar/m\omega}$. However, ψ_0 tends to zero asymptotically as x tends to $\pm \infty$, so that, according to the interpretation of $|\psi|^2$, there is a finite probability of finding the particle at very large distances from the origin. This is known as the *tunnel effect* and has an important application in the theory of radioactive decay.

The foregoing example demonstrates another point in which quantum mechanics differs from classical mechanics. According to the latter, the oscillator could have zero energy (when the particle is at rest at the origin), but this is not permitted in quantum mechanics—the *lowest* energy the oscillator can have is $\frac{1}{2}\hbar\omega$, which is called its *zero-point energy*.

Time Dependence. So far we have only considered the stationary states of a system and, furthermore, only of a *conservative* system, i.e., one whose energy is constant. However, if a system is subject to a disturbance which varies with time, for example, due to the passage of a charged particle or a light wave through it, there is a probability that the system will in a certain time make a transition from its initial stationary state to some other stationary state. In order to calculate such *transition probabilities*, it is necessary to consider explicitly the time dependence of wave functions, which we have so far neglected, and this entails the use of a modified Schrödinger equation.

Differentiation of the function $\Psi(x, t)$, Eq. (4), gives

$$E\Psi = i\hbar \frac{\partial \Psi}{\partial t} \quad (24)$$

which suggests that the energy E should in general be represented by the operator

$$E = i\hbar \frac{\partial}{\partial t} \quad (25)$$

If we convert the equation of energy [Eq. (12)] for a one-dimensional system into an operator, using Eqs. (13) and (25), and if we allow both

sides to operate upon a function $\Psi(x, t)$, we obtain

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \Psi}{\partial x^2} + V\Psi = i\hbar \frac{\partial \Psi}{\partial t} \quad (26)$$

[the derivative in Eq. (13) has been replaced by a partial derivative as we now have two variables, x and t]. This is known as the *time-dependent Schrödinger equation*; it applies even when the potential energy function V depends explicitly upon the time. If more than one dimension or several particles are involved, it is only necessary to use the appropriate quantum mechanical Hamiltonian operator on the left-hand side. It should be emphasized that this equation has *not* been rigorously derived, but only plausibly suggested. No such derivation exists, either for Eq. (26) or for Eq. (7)—their justification lies entirely in their successful application.

If we are, in fact, dealing with a conservative system, so that V does not depend explicitly upon t , it is easy to verify by substitution (and indeed this is obvious from the way in which the equation was derived) that Eq. (26) has solutions of the form

$$\Psi(x, t) = \psi(x) \exp(-iEt/\hbar) \quad (27)$$

where $\psi(x)$ is an eigenfunction of Eq. (7) and E is its corresponding eigenvalue. The function $\psi(x)$ is the time-independent wave function of a stationary state, and $\Psi(x, t)$ is the time-dependent wave function. Since it is not ψ but $|\psi|^2$ which has physical significance, and Eq. (27) shows that $|\Psi|^2 = |\psi|^2$, it is clear that so long as we are dealing with the stationary states of a conservative system, the time dependence of the wave function is of no importance.

Conclusion. The foregoing account has been confined to the elementary ideas of wave mechanics, due to de Broglie and Schrödinger. No mention has been made, for example, of the relativistic wave equation, due to Dirac, or of the important concept of electron spin. Further information on wave mechanics, its applications, and its relationship with the rest of quantum mechanics can be obtained from the article on MATRIX MECHANICS in this book and from a vast number of books of which those listed in the references are a very small sample.

S. RAIMES

References

- Born, M., "Atomic Physics," Sixth edition, London, Blackie and Son Limited, 1957.
- Merzbacher, E., "Quantum Mechanics," New York, John Wiley & Sons, 1961.
- Mott, N. F., and Sneddon, I. N., "Wave Mechanics and Its Applications," London, Oxford University Press, 1948; reprinted Dover, 1963.
- Raimes, S., "The Wave Mechanics of Electrons in Metals," Amsterdam, North-Holland Pub. Co., 1961.

Cross-references: MATRICES, MATRIX MECHANICS, PHOTOELECTRICITY, PHOTON, QUANTUM THEORY, SCHRÖDINGER EQUATION.

WAVE MOTION

Wave motion can be said to be the most common and the most important type of motion that we know. It is through wave motion that sounds come to our ears, light to our eyes, electromagnetic waves to our radios and television sets, and tidal waves and earthquakes to our cities. Wave motion can be defined as that mechanism by which energy is transported from a source to a distant receiver without the transfer of matter between the two points.

Waves can be classified according to the manner of their production, namely, a vibrating material object or, in the case of electromagnetic waves, sources such as electrical oscillations in an aerial. The wind blowing across water causes surface waves; a piezoelectric quartz crystal vibrating under an applied electric field generates underwater wave motion. Or, waves could be classified according to the medium in which they travel. The most useful classification, however, involves the direction of motion of the particles of a medium (or of an electric or magnetic field in the case of electromagnetic waves) relative to the direction in which the energy of the wave is itself propagated. Such a classification is useful because wave motions falling into the same class according to the selected criterion will have other similar properties.

Wave motion can be most easily understood if one considers first, as an example, wave motion in a horizontal, stretched string and then, by analogy, other types of wave motion. If one end of such a string is moved up and down, a rhythmic disturbance travels along the string. Each particle of the string moves up and down, while, at the same time, the wave motion moves along the length of the string. It is the state of the particles that advances, the medium as a whole returning to its initial condition after the disturbance has passed. Such a wave motion, one in which the vibratory motion of the medium is at right angles, or essentially at right angles, to the direction of propagation, is called transverse. Surface waves on liquids are transverse; so also are electromagnetic waves (x-rays, visible light, radio waves, and so forth), but here, since electromagnetic waves can travel in a vacuum, we must think of the electric and magnetic fields associated with such waves as changing in intensity in a direction at right angles to the direction of propagation.

Another type of wave motion, one termed longitudinal or compressional, can occur only in material media and has the particles of the medium moving forward and backward along the direction of propagation of the wave. Compressional waves are exemplified by sound waves in air, in which a volume in the path of the wave is alternately compressed and rarefied. These variations in pressure are very small. Even for the loudest sounds that an ear can tolerate, the pressure variations are of the order of 280 dyne/cm^2 (above and below atmospheric pressure of about $1\,000\,000 \text{ dyne/cm}^2$).

Yet another type of wave motion is the torsional wave, which can take place only in solids. Less frequently seen, this type can be demonstrated by a long helical spring supported on a flat surface. As one end of the spring is given a quick, momentary twist about the axis of the spring, a pulse travels down the spring.

Whatever the type of wave, certain useful definitions can be set forth and general statements made. Phase describes the relative position and direction of movement of a particle in its periodic motion as it participates in wave motion, of the relative intensity at a point of the electric field accompanying electromagnetic waves. Frequency is the number of complete vibrations performed per unit of time by a particle (or field) through which a wave passes. Period is the time required for one complete vibration of a particle participating in wave motion. Wavelength is the distance between any two points that are in phase on successive waves or pulses. The velocity of a wave is the product of the frequency and the wavelength. All waves except electromagnetic waves require a medium for their propagation.

Wave motions may vary in the energy they transport per unit of time. This property depends on the amplitude. The amplitude of a wave in a string, to take again an example, is the maximum displacement experienced by the particles of the string as they move from their equilibrium positions. The intensity of the wave is the power (energy per second) passing through a square centimeter perpendicular to the wave front, and it is related to the square of the amplitude. In the case of sound, intensity is related to loudness.

Two or more waves crossing one another's paths will not cause any change in the direction, frequency, or intensity of any of them. The displacement effects of two or more waves of the same kind passing through a medium are additive at any point; and at any moment, the displacement at a point is the vector sum of the separate displacements caused by the separate waves. Two transverse wave motions passing at right angles through a point in a medium cause a particle at that point to perform a path called a Lissajous figure, whose form depends upon the amplitudes and frequencies of the two waves and upon their phase relationship.

Wave motion also experiences the phenomena of absorption, reflection, refraction, interference, diffraction, beats, resonance, and polarization. However, longitudinal waves do not exhibit the property of polarization.

Complex waves can be analyzed into sets of simple waves according to the principles of Fourier analysis, where by "simple waves" are meant waves whose variations of displacement with time can be represented by sine curves.

Compressional waves require about 5 seconds to travel a mile in air, 1 second to travel a mile in water, and $1/3$ second to travel a mile in iron. Though varying in speed from material to material, low-frequency compressional waves travel with the same speed in a particular medium, i.e., they do not exhibit dispersion. Small

variations in speed sometimes found at high frequencies are due to relaxation phenomena.

Compressional waves in a fluid have a speed v that depends only on the density ρ and the adiabatic bulk modulus β of the medium according to the relation

$$v = \sqrt{\frac{\beta}{\rho}}$$

In solids, the speed of compressional waves is given by the relation

$$v = \sqrt{\frac{Y}{\rho}}$$

where Y is Young's Modulus. Thus it may be seen that a study of the propagation of waves in a medium gives important information about the medium.

Transverse waves on the surfaces of liquids do not travel with a fixed speed dependent only on the properties of the liquid. Their speed depends upon their wavelength and amplitude, the depth of the liquid, and whether the surface is confined, as in a canal. Surface tension waves on the surface of water have wavelengths less than 1.7 cm, while gravity waves on the surface of water have wavelengths greater than 1.7 cm. Ripples on water often move only 1 ft/sec. They have higher velocity as their wavelength becomes smaller. In contrast, for example, ocean waves measuring 800 feet from crest to crest have been found to travel 45 mph.

Transverse waves in strings (or wires) travel with a speed that depends only on the tension in the string T and the mass per unit length of the string m , according to the formula

$$v = \sqrt{T/m}$$

Electromagnetic waves of all frequencies travel with the same speed in a vacuum (2.9979×10^{10} cm/sec). In a particular medium, however, different frequencies (colors in the case of visible light) travel with different speeds. The speeds at particular frequencies also vary with the media.

The observed wavelength of a wave motion, whether it be longitudinal or transverse, depends on whether the source and the receiver are moving relative to one another, a phenomenon known as the Doppler effect. In the case of electromagnetic waves, the Doppler effect depends only on the relative velocity; in the case of a compressional wave, as for instance, sound, the magnitude of the effect depends not only on the relative velocity of the source and receiver but upon which is in motion with respect to the transmitting medium. In both cases the wavelength of the wave is increased if the source and receiver move away from one another and decreased if they move toward one another.

In some uses of wave motion, care must be taken to differentiate between "phase" and "group" velocity. The group velocity of a wave is usually the observed velocity, and the energy in a wave is transmitted with the group

velocity. The phase velocity refers to the speed of a point of fixed phase in the wave, as, for example, a crest in a water wave.

ROBERT T. LAGEMANN

Cross-references: ACOUSTICS, DOPPLER EFFECT, DYNAMICS, ELECTROMAGNETIC THEORY, HEARING, INTERFERENCE AND INTERFEROMETRY, LIGHT, MICROWAVE TRANSMISSION, MUSICAL SOUND, REFLECTION, REFRACTION, RESONANCE, SEISMOLOGY, SONAR, ULTRASONICS.

WAVEGUIDES

Fundamental Principles. A waveguide is a hollow pipe used as a transmission line. Its cross section is usually rectangular, square, or round, but irregular shapes are sometimes used for special applications. The electromagnetic wave in the waveguide can have an infinite variety of patterns, called modes, and, in general, these modes are of two kinds. In one set of modes, the electric vector is always transverse to the direction of propagation; in the other set, the magnetic vector is always transverse to this direction. The former are called TE or *transverse electric* modes, the latter are TM or *transverse magnetic*. Two subscripts are used to designate a particular mode. The first subscript (in either kind of mode) designates the number of half-wave variations of the electric field across the wide dimension of the waveguide, and the second designates the number of half-wave variations of the electric field across the narrower dimension. Thus, a $TE_{2,1}$ mode would designate a field pattern in which the electric field is always transverse to the propagation direction and in which the electric field has two half-wave variations across the wide dimension and one across the narrow.

For each mode of operation there is a cutoff wavelength, determined by the size and shape of the waveguide. For the $TE_{1,0}$ mode in a rectangular waveguide, the cutoff wavelength is twice the wide dimension. This means that when the signal wavelength is less than cutoff, the energy will propagate in the waveguide; but, when it is greater, the energy is attenuated exponentially. In the latter case, the frequency is too low to propagate through the waveguide and is said to be "below cutoff." Cutoff for other modes in rectangular guides is higher than for the $TE_{1,0}$. Consequently, for any frequency it is possible to choose dimensions so that only the $TE_{1,0}$ mode is above cutoff, and all other modes will be rapidly attenuated. The $TE_{1,0}$ mode is called the *dominant mode* and is the one most commonly used.

The attenuation through a waveguide is a function of the material of which it is fabricated. With suitable materials, the waveguide will have less attenuation at a particular frequency than a coaxial line of the same size. The losses in the waveguide are copper losses in the walls, and consequently the walls should be made of a

highly conducting material. At microwave frequencies the skin depth is only a few thousandths of an inch. Therefore, it is common practice to make the waveguide out of an inexpensive material such as brass, or a lightweight material, such as aluminum, and then to plate it with silver or copper for good conductivity. For prevention of oxidation, the conducting layer may be plated with a thin coating of rhodium. However, since rhodium has a higher resistivity than silver, this coating must be thinner than the skin depth of rhodium so that the current flows mainly in the silver layer.

Brass is the most common metal for waveguides because it is easily machined and easily soldered. For airborne applications where weight is a consideration, aluminum or magnesium is preferred. At very high frequencies, above 40 kMc, silver is frequently used, since the amount of metal used is small and costs less than the cost of a brass guide, silver plated.

Besides the simpler construction and lower loss, waveguides have higher power carrying capacity than coaxial cables. The waveguide is a completely shielded transmission line and may be bent and twisted with no radiation loss. However, whenever the waveguide is made to change direction, care must be taken to keep the cross section uniform or there will be a reflection from the discontinuity.

All discontinuities in waveguides are equivalent to lumped-circuit elements in conventional transmission lines. A screw inserted in the broad wall or a dent in this wall is a shunt capacity. A post across the waveguide is a shunt inductance. An iris across the waveguide with its edges parallel to the voltage vector is a shunt inductance. If its edges are parallel to the magnetic vector, the iris is a shunt capacity. A change in dimensions of the waveguide is a change in characteristic impedance, and a quarter-wave length of guide with new dimension is thus a quarter-wave transformer. All of these elements are used to match out the effects of more complicated mismatches.

In matching impedances or making quarter-wave transformers, the wavelength used is not the wavelength in free space, for the wave travels in the waveguide at a velocity apparently greater than that of light. Actually, this is the *phase velocity* of the wave and is the apparent speed of an unmodulated wave. The signal or modulation on the wave travels at a speed less than that of light called the *group velocity*. The product of the group velocity and the phase velocity is equal to the square of the speed of light. The phase velocity divided by the frequency is equal to the quantity called the *guide wavelength*, which is used for all matching calculations.

Waveguide Components. Two pieces of waveguide or two waveguide components may be joined together by means of flanges soldered on the ends of the guides. Flanges are of two types: (1) *cover flanges*, which are flat plane surfaces, and (2) *choke flanges*, which have a quarter-wave deep groove cut in them. This groove is also a

quarter-wave from the waveguide wall. The choke presents a short-circuit impedance between the two pieces of guide even if they are separated slightly or misaligned. It is thus possible to have a microwave coupling which is a short at radio frequency and an open circuit at direct current. By using a quarter-wave transformer, the reverse is also possible.

A choke joint is also used in a rotary or motional joint, where it is desired to move one waveguide with respect to another and at the same time to maintain electrical continuity. This goal is accomplished by physically separating the two pieces and using the choke to present an electrical short circuit at the microwave frequency.

In the rotary joint and in other microwave circuits, it is necessary to change from one form of waveguide to another or from waveguide to coaxial line. The problem in designing such adapters is basically one of matching impedances over a required frequency range. Coaxial line-to-waveguide adapters have been built with a voltage standing-wave ratio (VSWR) of 1.10 or better over a 30 per cent frequency band. Adapters from the dominant mode ($TE_{1,0}$) in rectangular guide to the dominant mode ($TE_{1,1}$) in round guide have been built with a VSWR of 1.05 over the same frequency range. In both cases the rectangular guide may be either perpendicular to, or in line with, the other transmission line.

The VSWR is measured by means of a *slotted line*. This is a section of waveguide, identical to that being tested, with a thin slot in the center of one broad wall, parallel to the direction of propagation. A probe is lightly coupled through the slot, and as it is moved along the slot, voltage maxima may be determined. This information can be used to calculate the impedance of the piece being tested.

In making the impedance measurements, the piece being tested must be properly terminated. That is, in order to measure only the discontinuity at the face of the slotted line, there must be no other discontinuities reflecting energy back to this point. The termination or *matched load* is usually a piece of waveguide containing an absorbent material which has been tapered to a point in the direction from which the energy is coming. Absorbent materials frequently used are resistance cards, carbon, plastics loaded with metal particles, sand, and wood. For very accurate measurements, the load material may be slid along the guide to determine whether it is properly matched. If a load is perfect, there will be no voltage variations observed at the slotted line probe as the load is moved in the guide.

A tee is a component consisting of a straight piece of waveguide with another piece fastened to it at right angles. The junction is open so that energy fed into any arm sees two possible paths at the junction. When the auxiliary arm is fastened to the broad wall of the main waveguide, the voltage or *E* vector in the side arm is perpendicular to the *E* vector in the main guide. Such a tee is called an *E-plane tee*. When properly matched, energy fed into the *E*-arm will divide

equally in the other two arms but 180° out of phase. This tee is also called a *series junction tee*. When the auxiliary arm is fastened to the narrow wall of the main waveguide, the magnetic or H vector in the side arm is perpendicular to the H vector in the main guide. This is an *H-plane tee* or *shunt junction tee*. When properly matched, energy fed into the H -arm will divide equally and in phase in the other two arms.

When the component has both an E-arm and an H-arm at the same point in the main line, it is called a *hybrid tee*. (The two main line terminals are now called side arms.) When properly matched, the tee has special properties and is called a *magic tee*. In this tee, energy fed into the E-arm or the H-arm divides equally in the other two arms and there is no coupling between the E- and H-arms. Energy fed into one side arm divides equally between the E and H arms, and there is no coupling to the other side arm.

When energy is fed into both side arms, the algebraic sum of the signal intensities appears at the H-arm, and the difference appears at the E-arm.

When two waveguides are joined together by two or more coupling paths, it is possible to choose dimensions such that energy in the first guide will be coupled a predetermined amount to the second and will travel in only one direction in the second guide. Then the output at one terminal in the second guide will be a measure of the power flowing in one direction only in the first guide. This component is called a *directional coupler*.

A piece of waveguide with two large discontinuities will have strong reflections between them. If the two discontinuities are spaced a multiple of half wavelengths apart, the reflections will reinforce each other and the section will be *resonant*. A resonant section or cavity is analogous to a fixed tuned circuit at lower frequencies. If the spacing of the discontinuities is adjustable, it is a variable tuned circuit. Since the frequency or resonance depends on the spacing, resonant cavities are used as wavemeters. The storage factor Q of the cavity depends upon the conductivity of the walls, the shape, the magnitude of the discontinuities, and the lightness of coupling. With silver plated walls and short-circuits, Q 's above 5000 have been obtained.

Attenuators and phase shifters are made by moving a piece of material into the waveguide so that it couples with the electric field. The motion can be calibrated and related to the amount of attenuation or phase shift. If the material is lossy, the component is an attenuator. If the material has low loss, but a dielectric constant more than unity, the velocity of propagation through it will be different from that in the empty guide, and phase shift results.

GERSHON J. WHEELER

WEAK INTERACTIONS

The weak interactions present a fascinating aspect of the problem of elementary particles. They appear at first to be totally unrelated to the other interactions that we know, nonetheless there have been discovered striking regularities among them, which can only be described in terms of symmetries which were originally ascribed to strong interactions alone. Without attempting to describe these, for which we must refer the reader to one of the books listed at the end of this article, we present here a brief history of the subject, which has been marked by many surprises.

To avoid the difficulties found in understanding processes of β -decay (see RADIOACTIVITY), in 1931, Pauli suggested that the emission of a β -particle from a nucleus is always accompanied by that of a spin-1/2 neutral particle of small mass, which takes account of energy and angular momentum conservation in β -decay. Calorimetric measurements of the average energy release in a β -decay transition agree quite well with the value calculated from direct measurement of the β -decay spectrum, indicating that the hypothetical particles, which were called neutrinos, must interact extremely weakly with matter, since they escape from the apparatus without giving up any measurable energy. A quantitative theory of β -decay, incorporating the NEUTRINO hypothesis was formulated in 1934 by Fermi, who wrote down the simplest relativistic expressions which would describe the basic β -transition: the transformation of a neutron into a proton accompanied by the creation of an electron and a neutrino. This theory has been strikingly successful in describing all aspects of β -decay—with one qualification to be described later—and is the one employed to this day. It could be anticipated from the weakness of the neutrino's interaction with matter that the coupling constant characteristic of β -decay would be small; it is in fact extremely small.* Adopting a simile of Lord Rutherford, one may say that from the point of view of the nucleus, β -decay practically never happens! To judge the scale, we note that lifetimes for β -decay are commonly several minutes whereas γ -transitions of similar energy occur in nanoseconds or less.

The success of Fermi's theory in accounting for the shapes of β -spectra and the dependence of the lifetimes on the available energy release could be regarded as indirect evidence for the existence of the neutrino; there were also other indications such as the occurrence of the predicted nuclear recoil after processes of orbital electron capture, a phenomenon predicted by the theory. Nonetheless, there was considerable satisfaction when the existence of the neutrino was directly demonstrated by Cowan and Reines in 1955 (for Reines' discussion see NEUTRINO). They used

Cross-references: CAPACITANCE, DIELECTRIC THEORY, INDUCTANCE, MICROWAVE TRANSMISSION, PROPAGATION OF ELECTROMAGNETIC WAVES.

* An average β -decay neutrino must pass through a thickness of matter of the order of 10^{20} g/cm² before suffering an interaction.

the intense neutrino flux arising from free neutron decays near a reactor to induce inverse β -decay, i.e., the conversion of a proton into a neutron with the emission of a positron, at roughly the expected rate. A similar experiment by Davis gave a negative result. In Davis' case, however, the reaction sought was the inverse of a β^+ -decay transition although the neutrinos available to him were the same as those used by Cowan and Reines, i.e., neutrinos arising from the β^- -decay of neutrons. The non-occurrence of the Davis reactions demonstrates that the neutrinos emitted in β^- -decay must be physically distinct from the neutrinos associated with β^+ -decay. This conclusion is supported by the absence of neutrinoless double β -decay, a process in which the neutrino from the β -decay of one nucleon could be reabsorbed within the same nucleus to induce the β -transition of a second nucleon.

Universal Fermi Interaction. The μ -mesons, or muons, (see ELEMENTARY PARTICLES) which were originally identified with Yukawa's mesons, were found to interact only very weakly with matter. In fact, it was noted that their absorption could be regarded as a process exactly similar to orbital electron capture, i.e., by an interaction just like that of β -decay, with even the same coupling constant! The decay of a muon, into an electron and two neutral particles, presumably neutrinos, was also accounted for by assuming a Fermi coupling of these four particles, again with a coupling constant of the same magnitude. This led to the hypothesis of a Universal Fermi Interaction between fermions, and after the discovery of "strange" particles (see STRONG INTERACTIONS), the natural extension of this hypothesis to include hyperons led to a qualitative understanding of all strange particle decays.*

The Universal Fermi Interaction does not, however, act between any set of four fermions. Interactions involving an electron or a muon always include a neutrino; furthermore, simple selection rules are found to operate for the weak interactions of strongly interacting particles. On the other hand, detailed experiments have revealed that the form of the Fermi Interaction is that of a current-current coupling, just as the electro-magnetic interaction of two systems may be regarded as the interaction between their currents. It has therefore been suggested that the Fermi interaction may arise as a result of the interaction between currents which generate a vector field similar to the electromagnetic field which mediates the weak interaction. To account for the short range of the Fermi interaction, the quanta of this field must be very massive, which may account for the fact that these particles have not been seen so far. The selection rules for weak interactions could be understood by assigning suitable properties to this field.

* The β -decay of hyperons occurs an order of magnitude more weakly than nuclear β -decay. To account for this in terms of an Universal Fermi Interaction, one must modify the definition of universality.

Fermi's theory predicts that the cross section for neutrino interactions increases quadratically with neutrino energy. This growth is not expected to continue after the neutrino wavelength becomes smaller than nucleon dimensions, but the cross section at the corresponding energy, about 1 GeV, is already sufficient for experiments using the high-energy neutrino beams from the decays of fast π - and K -mesons produced by high-energy accelerators. One of the first results of such experiments has been to demonstrate the striking fact that the neutrino associated with muon capture, which is the one predominantly produced in meson decays, is unable to produce electrons, proving that it is physically distinct from the neutrino of β -decay. High energy neutrino experiments can give us much more information about the structure of weak interactions than the limited number of decay processes which are available for study.

An aspect of weak interactions which we have not mentioned thus far, is that they possess fewer space-time symmetries than the strong interactions. Until Lee and Yang pointed out the possibility in 1956, it was not suspected that any interaction could distinguish between left-handed and right-handed coordinate systems.

It was then found that weak interactions discriminate between them very strongly; the question of how one system could be preferred over the other, was answered by Landau's principle of combined inversion. The apparent spatial asymmetry in a given experiment, is predicted to be exactly the mirror image of that to be expected in an experiment performed with the corresponding antiparticles; this prediction has been verified in several experiments. Fermi's theory, modified to allow for the nonconservation of parity, accounts almost perfectly for all weak interactions for which quantitative predictions can be made. A recent experiment,³ however, indicates a violation of the principle of invariance under combined inversion. Should this be upheld, it will demonstrate that weak interactions have more surprises in store.

P. KABIR

References

1. Fermi, E., "Elementary Particles," New Haven, Conn., Yale University Press, 1951.
2. Feynman, R. P., "The Theory of Fundamental Processes," New York, Benjamin, 1961.
3. Jackson, J. D., "The Physics of Elementary Particles," Princeton, N.J., Princeton University Press, 1958.
4. Kabir, P. K., Ed., "The Development of Weak Interaction Theory," New York, Gordon and Breach, 1963.
5. Christenson, J. H., Cronin, J. W., Fitch, V. L., and Turlay, R., *Phys. Rev. Letters*, 13, 138 (1964).

Cross-references: ELECTRON, ELEMENTARY PARTICLES, NEUTRINO, NEUTRON, PROTON, RADIOACTIVITY, STRONG INTERACTIONS.

WORK, POWER AND ENERGY

In the strict physical sense, work is performed only when a force is exerted on a body while the body moves at the same time in such a way that the force has a component in the direction of motion. The amount of work done during motion from point "a" to point "b" can be expressed by:

$$W = \int_a^b F \cos \theta \, ds$$

where F is the total force exerted and θ is the angle between the direction of F and the direction of the elemental displacement ds . In the cgs system the unit of work is the dyne-centimeter or erg, in the mks system the newton-meter or joule, and in the English system the unit of work is the foot-pound.

In rotational motion, the definition just given can be exactly applied, but it is often convenient to express the force as a torque and the motion as an angular displacement. The work done will be:

$$W = \int_a^b \tau \cos \theta \, d\omega$$

where in this case θ is always the angle between the torque τ expressed as a vector quantity and the elemental angular motion $d\omega$, also expressed as a vector. The units of work performed in angular motion will, of course, be the same as in the case of linear motion. Notice that the definition of work involves no time element.

Power is defined as the rate at which work is performed. The average power accomplished by an agent during a given period of time is equal to the total work performed by the agent during the period, divided by the length of the time interval. The instantaneous power can be expressed simply as

$$P = \frac{dW}{dt}$$

In the cgs system, power has the units of ergs per second; in the mks system, units of joules per second (or watts), and in the English system, units of foot-pounds per second. A common engineering unit is the horsepower, defined as 550 foot-pounds per second, or 33 000 foot-pounds per minute.

Energy may be defined as the capacity for performing work. This definition may be better understood when stated as: the energy is that which diminishes when work is done by an amount equal to the work so done. The units of energy are identical with the units of work, previously given.

Energy can exist in a variety of forms, some more recognizable as being capable of performing work than others. Forms in which the energy is not dependent upon mechanical motion are generally referred to as forms of potential energy. The most common example in this category is gravitational potential energy. A body near the earth's surface undergoes a change in potential energy when it is changed in elevation,

the amount being equal to the product of the weight of the body and the change in elevation.

Potential energy may also be stored in an elastic body such as a spring or a container of compressed gas. It may exist in the form of chemical potential energy, as measured by the amount of energy made available when given substances react chemically. Potential energy also exists in the nuclei of atoms and can be released by certain nuclear rearrangements.

Kinetic energy is the energy associated with mechanical motion of bodies. It is quantitatively equal to $\frac{1}{2}mv^2$ where m is the mass of a body moving with velocity v . In the case of rotational motion, the kinetic energy is more easily calculated using the expression $\frac{1}{2}I\omega^2$, where I is the moment of inertia of the body about its axis of rotation and ω is the angular velocity. Kinetic energy, like all forms of energy, is a scalar quantity. In a system made up of an assembly of particles, such as a given volume of gas, the total kinetic energy is equal to the sum of the kinetic energies of all the molecules contained in the volume. Calculation of the energy of such systems is very successfully treated theoretically on the basis of statistical averages.

Within a given system, energy may be transformed back and forth from one form to another, without changing the total energy in the system. A simple example is the pendulum, in which the energy is periodically converted from gravitational potential energy to kinetic energy and then back to gravitational potential energy. A similar situation, but on a submicroscopic scale, occurs in solid materials where the atoms are vibrating under the effect of interatomic rather than gravitational forces. As the temperature of a solid increases, the energy associated with the vibration of the atoms increases.

The example just given illustrates how, on a macroscopic scale, heat can be considered a form of energy. Regardless of the material involved, any amount of heat absorbed or released may be quantitatively expressed as an amount of energy. A gram-calorie of heat is equivalent to 4.19 joules, and in the English system a British thermal unit (Btu) is equivalent to 778 foot-pounds.

Potential energy is also present in electric and magnetic fields. The energy available in a region of electric field is equal to $E^2/8\pi$ per unit volume, where E is the electric field strength. Within a given volume, the total energy represented by the electric field is the integral of $E^2/8\pi$ over the volume. Similarly, the energy represented by a magnetic field may be independently calculated by integrating $H^2/8\pi$ over any given volume, where H represents the magnetic field strength. In the case of an electrically charged capacitor, the total energy in the electric field, and hence in the capacitor, can be shown to be $\frac{1}{2}CV^2$. Here C is the capacitance and V the electric potential to which the capacitor is charged. Similarly the total energy in the magnetic field associated with an inductor carrying an electric current is $\frac{1}{2}LI^2$, where L is the inductance and I the current.

Electromagnetic radiation is a combination of rapidly alternating electric and magnetic fields. Energy is associated with these fields and is exchanged between the electric and magnetic forms. This energy in a quantum of electromagnetic radiation, such as light or gamma radiation, can be represented in different ways, but is commonly expressed as $E = h\nu$. Here h is Planck's Constant and ν is the frequency of the radiation.

For particulate radiation or any very rapidly moving mass, the expression previously given for the kinetic energy, $\frac{1}{2}mv^2$, is not accurate when the velocity approaches that of the velocity of light. The theory of relativity requires a correction be made, and the exact kinetic energy, T , may be calculated in terms of the mass, m , of the body measured when at rest, and the speed of light in vacuum, c , as follows:

$$T = mc^2 \left[\left(1 - \frac{v^2}{c^2} \right)^{-\frac{1}{2}} - 1 \right]$$

Notice that this formula may also be written:

$$T = (m - m_0)c^2$$

where m is the variable quantity $m_0 \left(1 - \frac{v^2}{c^2} \right)^{-\frac{1}{2}}$.

This quantity represents the mass of the body, reducing to m_0 when v is zero, and approaching infinity as v approaches the speed of light.

This example illustrates another result of the theory of relativity, namely, the equivalence of mass and energy. Rewriting the last equation,

$$m = m_0 + \frac{T}{c^2}$$

The mass is seen to increase linearly with the kinetic energy of the body, the proportionality factor being c^{-2} . Indeed, even the rest mass, m_0 , represents an amount of energy equal to m_0c^2 . The total energy of a body of mass, m , can be generally given as:

$$E = mc^2 \text{ or } E = m_0c^2 + T$$

In dealing with radiation, whether particulate or electromagnetic, it is customary to express

energy in terms of electron volts. An electron volt is equal to the amount of work done when an electron moves through an electric field produced by a potential difference of one volt. One electron volt is equivalent to 1.60×10^{-12} erg. When charged particles such as electrons or protons are given kinetic energy by an accelerator, their kinetic energy is stated in terms of electron volts (eV) or million electron volts (MeV). In addition, any such particle will have a rest mass which can also be specified as an energy. For an electron,

$$m_0 = 9.11 \times 10^{-28} \text{ gram}$$

which is equivalent to 8.18×10^{-7} erg or 0.515 MeV.

A basic principle of physics is known as conservation of energy. This principle requires that within any closed system, the total energy must remain a constant. Energy can be changed from one form to another; but the total, so long as no energy is added to or lost from the system, must be constant. In the case of the swinging pendulum decreases in kinetic energy reappear as increases in potential energy and vice versa. Eventually, of course, the pendulum will stop due to the effect of frictional forces. At that time, all of the kinetic and gravitational potential energy will have been converted to heat.

In another example involving a radioactive atom, the total energy represented by the atom and the emitted radiation must be constant. If a gamma ray is emitted, the rest mass of the atom will be decreased by an amount equivalent to the sum of the energy of the gamma ray and the recoil kinetic energy of the atom, which will be very small. If a beta ray is emitted, the rest mass of the atom will be decreased by an amount equivalent to the sum of the rest mass of the emitted electron, the kinetic energy of the electron, and the recoil kinetic energy of the atom.

WILLIAM E. PARKINS

Cross-references: CONSERVATION LAWS AND SYMMETRY, DYNAMICS, MECHANICS, RELATIVITY, ROTATION—CIRCULAR MOTION, STATICS

X

X-RAY DIFFRACTION

X-rays are electromagnetic radiations, and like visible light, they can be diffracted (see OPTICS, PHYSICAL and DIFFRACTION). If a diffraction grating is used, the situation is as shown in Fig. 1. The points B, B', B'', ... along OX represent the lines of the grating seen end-on, and ABC, A'B'C', ... are typical incident and scattered

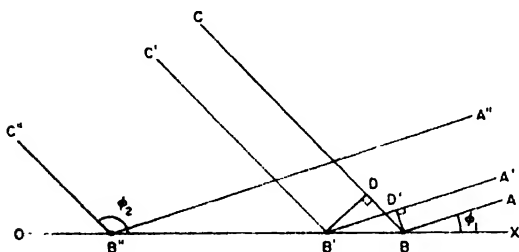


FIG. 1. Diffraction by a linear grating or line of scattering centres.

rays of a parallel incident beam. If D' is the foot of the perpendicular dropped from B on to A'B' and D is the foot of the perpendicular from B' to BC, the extra distance traveled by the ray A'B'C' is clearly D'B' - DB. This path difference can be expressed in terms of the grating space a , the angle of incidence ϕ_1 , and the angle of scattering ϕ_2 . From the triangle BB'D', D'B' = $a \cos \phi_1$, and from the triangle BB'D, DB = $a \cos \phi_2$. If there is to be appreciable intensity diffracted in the direction BC, the path difference must be an integral multiple of the wavelength λ , say $h\lambda$. Then

$$h\lambda = a(\cos \phi_1 + \cos \phi_2) \quad (1)$$

In order to obtain any appreciable scattering of x-rays from a grating it is necessary in practice to use very small angles of incidence, so that total external reflexion occurs. (The refractive index of matter for x-rays is very slightly less than unity, so that under suitable conditions they exhibit total external reflexion, whereas light exhibits total internal reflexion.) The points B, B', ... thus correspond to the centers of the smooth reflecting portions of the grating. From measurements of the angles of incidence and diffraction under such conditions, the absolute values of the x-ray wavelengths have been derived.

A geometrical representation of Eq. (1) is helpful in extending the theory of diffraction by a grating (an object with a one-dimensional variation of scattering power) to diffraction by a crystal (an object with a three-dimensional variation of scattering power). Equation (1) can be rewritten:

$$\frac{\cos \phi_1}{\lambda} + \frac{\cos \phi_2}{\lambda} = \frac{h}{a} \quad (2)$$

In Fig. 2, OP is a line of length $1/\lambda$ drawn parallel to BA, so that the angle XOP = ϕ_1 , and PQ is a line of the same length drawn parallel to BC, so that it makes an angle ϕ_2 with OX. The projection of OP on OX, OP', is clearly of length $\cos \phi_1/\lambda$, and the projection of PQ, P'Q', is of length

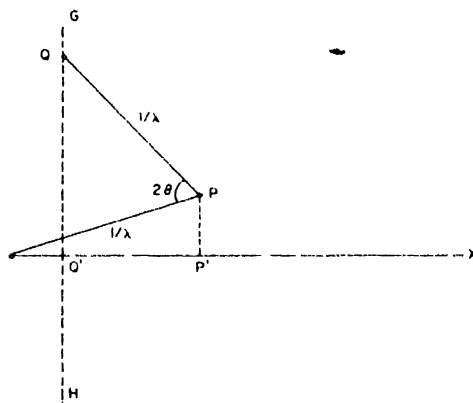


FIG. 2. Geometrical representation of the condition for strong reflection.

$\cos \phi_2/\lambda$. If Eq. (2) is satisfied, the length of $OQ' = OP' + P'Q'$ must be h/a . Clearly, fixing h is not sufficient to fix the angles ϕ_1 and ϕ_2 . If we imagine that the lines XO, OP and PQ are jointed together at O and P, Q is free to slide up and down the line GQ'H within wide limits without affecting the equality of Eq. (2), and so giving considerable freedom to the directions of incident and diffracted rays having a constant path difference $h\lambda$. Different values of h lead to different lines GH, G'H', ... spaced at integral multiples

of the distance $1/a$ along OX. All possible mutual relationships between the incident and strongly diffracted rays are represented by those portions of the set of parallel lines of spacing $1/a$ that lie within a circle with radius $2/\lambda$ and center O.

The exterior angle between AB and BC is generally denoted by 2θ . From Fig. 1 it is clear that

$$2\theta = \phi_2 - \phi_1 \quad (3)$$

and from Fig. 2

$$\begin{aligned} OQ &= OP \sin \theta + PQ \sin \theta \\ &= (2 \sin \theta)/\lambda \end{aligned} \quad (4)$$

With a little reinterpretation, Figs. 1 and 2 can be used to derive the condition for strong diffraction by a crystal. A crystal is a three-dimensionally periodic object, with a certain arrangement of atoms repeating indefinitely at intervals of a in one direction, b in another, and c in a third. The intervals a , b , c are not generally equal to each other, nor necessarily at right angles to each other. They define a parallelepiped called the *unit cell* of the crystal. For determining the geometrical conditions for the existence of a strong diffracted beam, the manner in which the atoms are arranged in the cell is not important, and for the moment we may think of them as forming a single scattering center at the corner of the cell. In Fig. 1, the points B, B', ... now represent a row of scattering centres parallel to a , and ABC, A'B'C', ... represent a sequence of incident and diffracted rays. The difference is that for a ruled grating one can expect constructive interference between the various scattered rays only if OX, AB, BC, A'B', B'C', ... are coplanar, whereas an arrangement of atoms can scatter out of the plane of OX and AB, and hence BC need not lie in this plane. The condition for reinforcement, however, remains

$$D'B' - DB = h\lambda \quad (5)$$

and its geometrical representation in Fig. 2 is changed only in that Q does not necessarily lie in the plane of OX and OP. With this relaxation, the condition for reinforcement is fulfilled if Q lies anywhere on a plane perpendicular to a and distance h/a from O. Different values of h lead to a set of parallel planes, and as far as the repetition of scattering centers in the direction of a is concerned, reinforcement will occur if Q lies anywhere on any of the planes within a sphere with center O and radius $2/\lambda$.

It is, however, necessary to satisfy similar conditions for the repetition of scattering centers in the directions of b and c . These conditions are exactly analogous: reinforcement for repetition along b will occur only if Q lies in one of a set of planes that are perpendicular to b and spaced $1/b$ apart, and reinforcement for repetition along c will occur only if Q lies in one of a set of planes perpendicular to c and spaced $1/c$ apart. There can be a strong beam diffracted by the crystal only if Q satisfies these three conditions

simultaneously. Two sets of planes intersect in a set of parallel lines, and the third set of planes will intersect these lines in a lattice of points. This is called the lattice reciprocal to the crystal lattice, or the *reciprocal lattice* for short. A strong diffracted beam can occur, then, only if Q coincides with one of the points of the reciprocal lattice, and only those points within a sphere of radius $2/\lambda$ are possible, since the maximum length of OQ is $2/\lambda$. This sphere is called the *limiting sphere*.

The repeat distances in the reciprocal lattice are usually called a^* , b^* , c^* . They have the same general directions as a , b , c , but are actually parallel to them only in those crystal systems that have their axes at right angles. There are a number of "reciprocal" relations between the two sets of axes, some obvious from the way in which the reciprocal lattice has been constructed, others obtainable only by analysis. For further details the article on CRYSTALLOGRAPHY should be consulted.

For a given wavelength λ , a given direction of incidence PO, and a fixed position of the crystal, the loci of possible positions of Q is a sphere with center P and radius $1/\lambda$. This sphere is called the *sphere of reflexion*, since strong diffraction can take place only if one of the points of the reciprocal lattice lies in or passes through its surface. With all parameters fixed, this is an unlikely coincidence, and in the practical study of x-ray diffraction, provision must be made for variation of either λ (the Laue method), or the orientation of the crystal (most types of single-crystal and crystal-powder cameras and diffractometers), or the direction of the incident x-rays (some special-purpose techniques). Space does not permit discussion of the practical details.

In the preceding treatment diffraction has been considered from the viewpoint of adding contributions scattered by centers arranged in a regular lattice. It can also be considered as reflexion from sets of parallel planes of atoms, the rays reflected by successive planes in a set being added. It can be shown that the connection between the two views is that (1) the line OQ joining the origin of the reciprocal lattice to the point hkl is perpendicular to the corresponding set of reflecting planes, and (2) its length is the reciprocal of their spacing d . From the second viewpoint, then, Eq. (4) may be rewritten

$$\lambda = 2d \sin \theta \quad (6)$$

a relation known as *Bragg's law*.

The arrangement of atoms within the unit cell influences the intensity of the various orders of diffraction hkl , just as the shape of the ruling influences the intensities of the different orders produced by an optical diffraction grating. It is easy to see that x-rays diffracted by an atom at the position xa , $y b$, $z c$ within the unit cell travel a shorter distance than those diffracted at the hypothetical scattering center at the origin, and that the corresponding phase difference is $2\pi(hx + ky + lz)$. If the cell contains n atoms whose scattering

factors (the ratio of the actual scattered amplitude to the amplitude scattered by a free electron under the same conditions) are f_1, f_2, \dots, f_n , the total amplitude scattered by the cell will thus have an in-phase component of

$$A = f_1 \cos 2\pi(hx_1 + ky_1 + lz_1) + \dots + f_n \cos 2\pi(hx_n + ky_n + lz_n) \quad (7)$$

and an in-quadrature component of

$$B = f_1 \sin 2\pi(hx_1 + ky_1 + lz_1) + \dots + f_n \sin 2\pi(hx_n + ky_n + lz_n) \quad (8)$$

the scattering by an atom at the origin of the cell being taken as the reference phase. The total intensity of scattering is thus proportional to

$$F^2 = A^2 + B^2 \quad (9)$$

The quantity F is called the *structure factor*; it is sometimes convenient to regard it as a complex quantity $F = A + iB$. It can vary in magnitude from zero to $f_1 + f_2 + \dots + f_n$, with a root-mean-square value of $\sqrt{(f_1^2 + \dots + f_n^2)}$. The total intensity of the order hkl is proportional to NF^2 , where N is the number of unit cells in the crystal.

The maximum intensity of diffraction occurs when the point Q , defined above, coincides with one of the points hkl of the reciprocal lattice. As the point Q moves away from its ideal position, the intensity of diffraction does not drop instantaneously to zero, but falls off gradually in a manner depending on the size and perfection of the crystal. The study of the manner in which the falling off occurs is a fascinating exercise both experimentally and theoretically, but space permits only the quotation of a few qualitative results. The main varieties of imperfection considered are (1) the crystal is too small to give sharp diffraction maxima, (2) there are "mistakes" in the arrangement of atoms in the sequence of unit cells, and (3) the crystal is twisted, bent, or affected by dislocations. In case (1) the regions round each reciprocal-lattice point are identical except for variations in F , and extend in any direction to a distance that is roughly the reciprocal of the thickness of the crystal in the corresponding direction. In case (2) the size and shape of the regions may vary in a complex manner with hkl , but they show no general increase in size with increasing distance from the origin of reciprocal space. In case (3) the size of the regions shows a general increase in direct proportion to their distance from the origin of reciprocal space, as well as varying with hkl .

A. J. C. WILSON

References

- Guinier, A., "X-ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies," San Francisco, Freeman, 1963.
- Hosemann, R., and Bagchi, S. N., "Direct Analysis of Diffraction by Matter," Amsterdam, North Holland, 1962.
- James, R. W., "The Optical Principles of the Diffraction of X-rays," London, Bell, 1948.
- Wilson, A. J. C., "X-ray Optics," Second edition, New York, John Wiley & Sons, London, Methuen, 1962.
- Zachariasen, W. H., "Theory of X-ray Diffraction in Crystals," New York, John Wiley & Sons, 1945.

Cross-references: CRYSTALLOGRAPHY; DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS; ELECTRON DIFFRACTION; NEUTRON DIFFRACTION; OPTICS, PHYSICAL.

X-RAYS

Although x-ray (or Roentgen ray) science was 70 years old on November 8, 1965, it maintains a youthful vigor and an ever-accelerating pace in new developments unsurpassed in all other areas of pure and applied science—physical, chemical, biological, medical, industrial or engineering. On that date in 1895, at the University of Wurzburg, Wilhelm Conrad Röntgen observed the fluorescence of a barium platinoeyanide screen, excited by the emission from a Crookes cathode-ray tube of a "new kind of ray", penetrating through space invisible, unaffected by electric or magnetic fields, differentially absorbed in passing through matter of varying composition, density or thickness, ionizing gases and affecting electrical properties of liquids and solids, affecting the photographic plate, and producing other chemical and biological reactions. The obvious similarities with light led to the crucial tests of established wave optics: polarization, diffraction, reflection and refraction. With limited experimental facilities, Röntgen and his contemporaries could find no evidence of any of these, hence the designation x (unknown) of x-rays generated by the stoppage at anode targets of cathode rays, identified by J. J. Thomson in 1897 as electrons, accelerated by a potential of a few thousand volts in evacuated tubes.

In 1906 Barkla found evidence in scattering experiments that x-rays could be polarized or constrained to vibrate in one direction perpendicular to the direction of beam propagation. The really convincing proof of the nature of x-rays, placing them in the long range of electromagnetic waves with visible, ultraviolet and infrared light, radio and Hertzian waves, and eventually with gamma rays from radioactive sources, was the discovery of diffraction of collimated beams by crystals acting as three-dimensional gratings (simultaneously proving the regularity of crystalline architecture). The beautiful diffraction pattern of a zinc sulfide crystal photographed in 1912 by von Laue and associates was interpreted soon thereafter by the simple Bragg law: $n\lambda = 2d \sin \theta$ where n is an integer—the order of diffraction, λ is the wavelength of the x-ray beam, d is the interplanar grating spacing of a crystal, and θ is the

angle of incidence (2θ is the angle of diffraction). Thus was founded the science of x-ray spectrometry (λ unknown, d known) and of x-ray diffractometry or crystallography (λ known, d unknown), the experimental measurement in both cases being the angle θ or 2θ .

In 1921 Compton and associates proved that x-rays are reflected from mirrors at small grazing angles of incidence, and are refracted very slightly in the opposite direction from light, and thereby diffracted by ruled gratings. Thus the wave properties of x-rays were finally convincingly demonstrated, but at the same time, the dual nature of the radiation as both waves and corpuscles or photons or quanta became clearly evident. The latter certainly were involved in the photoelectric effect: (1) by the short wavelength limit of the general radiation spectrum sharply defined and represented by $Ve = hc/\lambda_0$ (law of Duane and Hunt, 1915), where V is the voltage applied to the x-ray tube, e is the electronic charge, h is the Planck action constant for quanta, and c is the velocity of light, and (2) by the COMPTON EFFECT, 1922, or an increase in wavelength of scattered radiation as a function of angle, the lost energy appearing as electron recoil.

From 1895 to 1912, the nature of x-rays remained mysterious since there was no method of dispersing the radiation from the target of a tube into a spectrum such as characterized beams of light after refraction in prisms or diffraction by ruled gratings. However, absorption measurements indicated important variations in "quality" — "hard" and "soft" rays — depending upon the target element and the voltage, and the distribution of emitted energy in groups for which Barkla in 1905 devised the nomenclature K, L, M, N, O, P series, by analogy with optical spectral series. The way was opened for the design of the Bragg crystal spectrometer with the discovery of crystal diffraction and the derivation of the Bragg law. With the spectrometer, whose principles still prevail in the automatic high-precision instruments of today, x-ray beams were resolved into spectra of sharp lines in series superposed on a continuous background (the continuous, general or "white" radiation), whose wavelengths were characteristic for each

element. There were also characteristic absorption discontinuities in spectra. Thus Bragg, Moseley, Siegbahn and others began the accumulation of the lists of characteristic wavelengths. Moseley was the first to recognize the essential similarity and simplicity of the K-series, for example, with the same 4 principal lines with wavelengths for each which varied continuously step by step in proceeding from one atomic species to the next in the periodic table. In the immediate intensive effort to account for the origin of x-ray spectra in terms of atomic structure, it was clear that the atomic number [now 1 (hydrogen) to 103 (lawrencium)] is the number of external electrons or the net positive charge on the nucleus of the atom. In this way, the theories of atomic structure arose. First there was the Bohr theory in 1914 of miniature solar systems involving revolutionary concepts introducing the quantum theory of Planck and Einstein. Then followed the gradual change from the mechanical model to the present vector or quantum mechanical model which can only be described mathematically, but with virtually no change in the interpretation of x-ray spectra. The tables of readily available wavelengths are measures of the energies of orbitals in the outer electronic shells of atoms around the nuclei. The original uses of the spectrometer for these physical aspects are fairly well complete and have given way principally to chemical analysis. Spectrometry is only one of at least 25 branches of this radiation science, depending upon techniques, information provided, and modern applications. The attempt is made here for the first time to present a condensed tabulation of these 25 methods (see Table 1). Details in all cases may be found by reference to the companion volume of this Encyclopedia—"The Encyclopedia of X-Rays and Gamma rays," edited by the writer and published by Reinhold Publishing Corp. in 1963.

GEORGE L. CLARK

Cross-references: DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS, ELECTRON OPTICS, REFRACTION, X-RAY DIFFRACTION.

TABLE I

Method	Instrumental Technique	Information and Uses	Recent Advances
(1) Emission spectrometry	Crystal spectrometer analysis of fluorescent characteristic rays generated in specimen by primary x-rays; photographic film; Geiger, proportional or scintillation counters	Qualitative and quantitative chemical analysis for elements above Cr(6)	1. Automatic, programmed apparatus with discrimination and pulse height analysis. 2. Trace analysis especially in petroleum. 3. Non-dispersive (no crystal) spectrometry greatly extended. 4. Multiple automatic analysis of up to 24 elements. 5. On-stream control of processes (Mg in cement, etc.). 6. Quantitative theory and practice of corrections for matrix effects. 7. Extensions to very long wave-lengths—carbon, 45Å, boron 67Å (Henke)

Method	Instrumental Technique	Information and Uses	Recent Advances
(2) Electron probe microanalyzer	Electron beam collimated by magnetic lenses generates primary rays from areas of $1\mu^2$ (vol $1\mu^3$), analyzed by sensitive spectrometer	Analysis of very small areas or single grains for diffusion, segregation, microscopic processes in alloys, minerals, etc.	1. Several commercial instruments up to \$100 000 on market. 2. Scanning of specimen by electron beam with 1 crystal setting to show distribution of each element as magnified image 3. Accelerating value in biology and medicine. 4. Amazing success in detecting cause of failure in electronic components for guidance systems, semiconductors, etc.
(3) Absorption edge spectrometry	Crystal spectrometer analysis of characteristic absorption by elements in screen from continuous primary spectrum (1K, 3L, 5M edges)	As in (1); long considered relatively insensitive, independent of state or valence of elements	1. Greatly sensitized, automatic apparatus to measure edge height, free from interferences in emission spectra. 2. Highly successful at Oak Ridge for U, Th, Zr, Nb, Mo, Hf, W in presence of variety of matrix elements
(4) Fine Structure of absorption edge	High-resolution spectrometer used as in (3)	K-edge found to have a fine structure dependent on valence of absorbing element	Spectacular success in determining valence state of elements in catalysts such as NaMnO ₂ , on silica unchanged Mn ²⁺ , on charcoal Mn ¹⁺ (Herbstein, South Africa)
(5) Scattering	Fluorescence spectrometer as in (1); sample gas, liquid, solution or solid	In addition to emission lines coherent scattered rays (WLa ₁ from W-target) and incoherent (Compton) rays at larger angle; former detects high atomic number elements in low atomic number matrix	Ratio of intensities, coherent/incoherent, powerful analysis for carbon and hydrogen in organic compounds calibrated with known hydrocarbons and corrections for other elements, more rapid, simpler, equally dependable compared with micro-combustions
(6) Low-angle (Gunnier) scattering	Monochromatic beam falling on certain colloidal samples scattered at very small angles from direct beam (distinguished from small-angle diffraction), as measure of electron density inhomogeneity	Colloids or heterogeneous systems of clumps, holes, voids, pores with discontinuity from 10 to 10 000 Å, measured for radius of gyration, <i>R</i> , related to particle size	1. Until recently purely academic but now widely used with proteins, viruses, catalysts, carbon blacks, hardening and precipitation in alloys, lattice deformations. 2. Invaluable for study of critical phenomena (liquid-vapor) 3. Best method for voids in fibers, vitreous silicates, etc
(7) Absorptometry	Measurement of attenuation of x-ray beam passing through matter with intensity before and after; polychromatic and monochromatic beams	Chemical analysis of gases, liquids, solutions, solids, with proper calibration; true densities, porosities; coating, plating, insulation thickness	1. Automatic double-beam photometers with chopper, and multiplier phototubes with vastly improved phosphors and intensification, often linked with servomechanisms. 2. Automatic gauging on-stream in metal rolling mills 3. Automatic densitometry of static and fluid beds, and porosity of coke etc. 4. <i>In vivo</i> measurement of bone density, iodine in thyroid (dichromatic)
(8) Radiography	Registration on film of differential absorption of beam passing through specimen of varying composition density ("shadow-graphs")	Medical diagnosis; industrial testing of internal soundness (castings, forgings), correct fabrication, etc.	1. Manifold increase in nondestructive testing in art museums, especially paintings for authenticity, repair, hidden images. 2. Wide advance of supervoltage radiography of thick sections (betatrons, Van de Graaff and linear accelerators.) 3. New use of color film, polaroid 10-second developing film, Xerox images, and far more frequent enlargement of images
(9) Flash and cineradiography	Instantaneous and motion picture sequences with extremely intense beams	Recording of events radiographically in times as short as 1 nsec (10^{-9} sec)	1. Field emission (sharp-pointed cold cathode) tubes with currents of hundreds of amperes in pulses and intensities in millions of roentgens. 2. Used in ballistics, explosions. 3. Biological use illustrated by experiments on small animals strapped in car shot from bow at 30G acceleration, showing amazing displacement of internal organs (significant for human astronauts)

Method	Instrumental Technique	Information and Uses	Recent Advances
(10) Fluoroscopy	Shadows of gross structures registered on fluorescent screen	Medical observation of internal motions and course of surgery; motions of machinery. Always handicapped by weak intensity and exposure danger to patients; shoe-fitting prohibited in many states by law	1. Technique revolutionized by electronic intensifiers in use of very weak primary rays free from exposure danger. 2. Growing use of radioisotope x-ray sources hitherto far too weak, for mobility in field, simplicity, high penetration of high-energy rays
(11) Tomoscopy	X-ray source and intensifier on each side of object, projecting image on fluorescent screen, moved synchronously, parallel and in opposite directions	Produces image of a single plane in interior of object; serioscopy a series of planar images through object by readjustment	Entirely new because of efficient intensifiers, to localize exactly any feature obscured by 3 dimensions superposed in usual fluoroscopic image
(12) Contact microradiography	Beam of monochromatic or very soft x-rays pass through small specimen in contact with photographic emulsion, and image enlarged, usually 100 \times , with interference from graininess	One type of x-ray microscopy known for 50 years but used only recently, useful for any kind of specimen with heterogeneous absorbing power	1. New fine-grained emulsions permitting higher magnifications up to 800 \times . 2. Grainless media such as polymer sheets cross-linked by exposure to change solubility and form relief of image shadow-cast and examined by electron microscope. 3. Single-cell images by use of very soft rays such as C-K α at 45Å (Henke) 4. Microangiography (blood vessels with opaque stains) developed at Mayo Clinic as accepted diagnostic method 5. Stereomicroradiography with ingenious apparatus for 3-dimensional view.
(13) Microfluoroscopy	Absorption image on fluorescent screen microscopically examined and photographed	Useful for observation of changes or even cellular motion at higher magnifications than with film from much smaller grain size than silver halides in film	Entirely new technique by Pattee (Stanford)
(14) Projection microscopy	X-rays from point source generated by electrons collimated by magnetic lenses registered as sharp images on film or screen at various distances with inherent enlargement without penumbra or limitation of graininess	Same as (12) and (13) but much more direct, rapid and higher magnifications	Removes all interference from graininess of film since enlarged image formed directly; dramatic micrographs of freeze-dried biological specimens now being published
(15) Reflection microscopy and telescropy	Original Kirkpatrick microscope with true optics of total reflection of x-rays from cylindrical or elliptical lenses in multiple	Theory and practice at stage of rapid development for microscope and for telescopic enlargement of solar and celestial x-rays	See column at left
(16) Radiation chemistry, radiolysis	Irradiation of all types of materials from intense source and evaluation of chemical change from absorbed dose in rads (100 ergs/g)	Radiation-induced chemical reactions by formation of intermediate excited species, ions, free radicals; thus many dosimeters and potential radiation processes, and aid in accounting for biological effects	1. Immeasurably great recent progress in mechanisms and in manufacturing - ethyl bromide (Dow), phenol from benzene and water (Russia), cross-linking of polyethylene, etc., for insulation, removal of mercaptans from petroleum, graft polymerization, rubber vulcanization, synthesis of organic compounds with CO ₂ , producing NF ₃ oxidizer for rocket fuels. 2. Discovery of hydrated electron as essential intermediate in aqueous systems by E. J. Hart. 3. Acceptance of G (M. Burton) for reporting yields; number of molecules reacting or produced per 100 eV absorbed radiation energy

Method	Instrumental Technique	Information and Uses	Recent Advances
(17) Radiation physics	Same as (16)	Directed to study of "damage" in solids by excitation and ionization of electrons, atomic displacements or defects in lattices with effects on mechanical, thermal, electrical, magnetic, optical, color properties	1. Now an essential tool of solid-state physics, in research and production of semiconductors-transistors, rectifiers, diodes, infrared detectors, solar batteries. 2 Just announced in advertisements irradiation of LiF crystals to permit cleavage for use in many instruments.
(18) Radiation biology	Same as (16) and (17) on living systems	Effects of radiation exposure on all organisms, especially sterilization, potato sprout inhibition, intensive research since "fallout" on chemical, physical and biological changes in cell constituents on up to whole body	1. Food preservation without refrigeration by bacterial sterilization, a major project of the Army and AEC. Bacon approved in 1963 by FDA to be followed by meats, chicken, fish, fruits, vegetables. 2. Soybean seeds surviving a million roentgen x-ray dose with improved yields, unexpectedly due to lethal effect on 4 million bacteria per seed with release of pyrogenic protectant of plant cells
(19) Radiation genetics	Same as (18) with special concern with reproductive cells	To produce and study mutations some it is hoped useful, by effects on DNA, chromosomes and genes	1. More than ever before a search for favorable mutations in crop plants, flowers, etc. 2. Intensive research related to cancer cells, nature of DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) in terms of variations in amino acid sequences
(20) Radiation therapy	Irradiation of cancer and other neoplasms in controlled doses, massive and fractional	Slow, step-by-step, upward climb with mounting hopes, with best results from supervoltage (1-12 MeV) rays	Increasing use of isotopic sources of x-rays, as well as γ -rays, and skillful combination with new chemotherapeutic agents, surgery and transplantations
(21) Radiation protection	Special case of (16) in test of chemicals as protective agents against x-rays in chemical and biological systems,	Preferential absorption or energy, destruction of free radicals or other intermediates from ionizing rays, by wide range of compounds	Among thousands of compounds tested, sulfur-containing cysteine, cysteamine and ALI (ammonioethylisothiuronium bromide hydrobromide, now on market in tablet form) most extensively studied and among most effective in protecting living organisms against deleterious radiation effects when inbibed before exposure, toxicity of drugs serious factor and shielding from radiation still surest protection
(22) Single-crystal analysis	Diffraction of x-rays by crystals discovered by von Laue in 1912, Laue stationary film and sample, rotating crystal, and moving-film methods for patterns	Ultimate crystal and molecular structures of increasingly complex compounds, derived from patterns, penicillin analysed to synthesis; control of cutting of quartz crystals for electronics, and identification of gems among industrial uses	1. Vast improvements in automatic apparatus and intensity measurements, in mathematical theory and interpretation such as Fourier, Patterson, Hosemann functions, fold integrals, ingenious handling of inherent phase problem, in aids such as computers, optical analog devices and summators, reciprocal lattice concept, molecular transforms, stochastic methods. 2. Successful analysis of many complex biologically important compounds such as all types of fibrous proteins. 3. Quantitative detection and evaluation of thermal oscillations producing anomalous interferences, coefficients of expansion, mosaic structure and imperfections in single crystal
(23) Powder diffraction analysis	Patterns of random powders and aggregates in many kinds of cameras yielding line patterns, each a "fingerprint" of a given element or compound, many listed in ASTM card index	Identification of crystalline material, quantitative analysis of mixtures, solid solutions, reactions, etc.; one of the most widely used instrumental methods of analysis	1. Automatic recording programmed diffractometers of great value in time saving, accuracy and interpretation. 2. New microcameras and complete apparatus for patterns at low and high temperatures and pressures, essential in phase studies. 3. Exceptional advance in precision measurements of lattice parameters indicative of lattice perfection, atomic weights, ideal densities, polymorphic equilibria, etc. (Straumanis, Beu <i>et al.</i>). 4. Radial distribution analysis of amorphous materials

Method	Instrumental Technique	Information and Uses	Recent Advances
(24) Texture analysis	Diffraction patterns by usual techniques of (23) plus special apparatus such as pole-figure goniometers, back-reflection, etc.	Determination whether crystalline or amorphous, per cent of each, preferred orientation, grain sizes in microscopic and colloidal ranges, strains and defects.	The essence of solid-state science and the area of greatest progress in theory and practice in study of materials in terms of behavior, beyond actual lattice structure, and changes due to environmental effects; covered in massive new treatise published by Springer, Germany
(25) Topography	Special textural analysis, technique due to Lang (Bristol) usually with specimen and film traversed back and forth together in x-ray beam	Since strained and imperfect crystal area diffract more strongly than perfect, dislocations beautifully detected and identified topographically	One of the great new techniques of utmost practical value in disclosing how and why materials fail, how crystals grow, what physical and mechanical measurements of tensile strength and deformation mean.

Z

ZEEMAN AND STARK EFFECTS*

In 1896 Zeeman observed that the yellow lines of a sodium flame were considerably broadened when it was placed in a strong magnetic field. Using his electron theory of matter and radiation, Lorentz showed that the light should be circularly polarized when viewed along the field (σ components), and linearly polarized along (π components) and perpendicular to the field direction (σ components), when viewed transverse to the magnetic field. This was verified by Zeeman using a Nicol prism to analyze the polarization.

A magnetic field exerts a force $-e/c[\mathbf{v} \times \mathbf{H}]$ on an electron with a velocity \mathbf{v} . Hence, a field H_z does not affect motion in the z direction, but electrons moving in counterclockwise directions in the $x-y$ plane are speeded up by the magnetic field, while those moving in a clockwise direction are slowed down giving frequency changes

$$\Delta\nu = \pm eH/4\pi m_e c = 1.4H \times 10^6 \text{cps}$$

(e is the electronic charge in electrostatic units, H is the magnetic field in oersteds, m_e is the mass of the electron and c is the velocity of light). Expressed in wave numbers

$$\Delta\tilde{\nu} = eH/4\pi m_e c^2 = 4.6688 \times 10^{-5} H \text{cm}^{-1}$$

which is called a Lorentz unit. Figure 1 shows the normal Zeeman effect, or Lorentz triplet, obtained with this theory. Such triplets were soon observed in cadmium and zinc by Preston, and provided clear indication that the emission of light by atoms involved the motion of negatively charged electrons. Most Zeeman patterns are anomalous, however, and show more than three components (see Fig. 1). Such complex patterns require a quantum mechanical interpretation, including the concept of electron spin. They depend only on the angular momentum quantum numbers of the terms in the spectral series (Preston's rule), and all known patterns are rational multiples of the normal triplet separation (Runge's rule).

According to quantum theory, the angular momentum, $P_l = \hbar/2\pi$, of the electron is quantized and is represented by a vector \mathbf{l} normal to the orbit. The magnetic moment of the electron is then $\mu_l = -le/2m_e c$. Magnetic field effects are then equivalent to a precession of \mathbf{l} around \mathbf{H} at

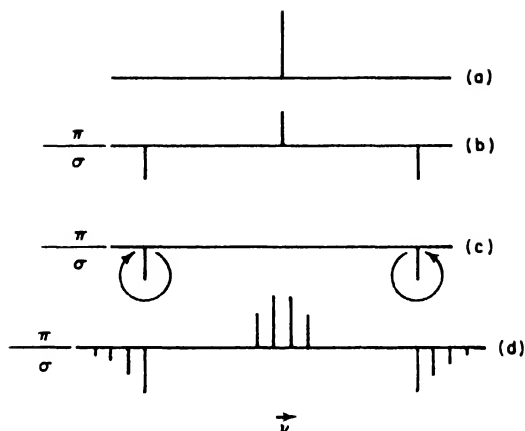


FIG. 1. Zeeman effect on spectral line: (a) No magnetic field—unshifted line. (b) Normal Zeeman effect. Transverse magnetic field, π and σ components of normal Lorentz triplet. (c) Normal Zeeman effect viewed along magnetic field. Circularly polarized σ components. (d) Anomalous Zeeman effect of the transition $2d_{5/2}-2P_{3/2}$.

the Larmor frequency $\nu_L = eH/4\pi m_e c$. Only discrete angles between \mathbf{l} and \mathbf{H} are allowed given by the projection m of \mathbf{l} on \mathbf{H} , where m takes values $-l$ to $+l$. The energy splitting of the l term is thus $m\hbar eH/4\pi m_e c$, and the selection rules $\Delta m = 0, \pm 1$ apply for allowed transitions between different l terms. Corresponding frequency shifts are 0 and $\pm eH/4\pi m_e c$, which again gives the normal Zeeman effect. Landé used such vector models and introduced empirically the g_l factor for each l term, with energy separations $\Delta E_l = m\hbar g_l eH/4\pi m_e c$. He was able to deduce the appropriate g_l factors and explain the Zeeman patterns to a remarkable degree.

The spin of the electron, postulated by Uhlenbeck and Goudsmit, allowed a more satisfactory deduction of these g factors. To account for the doublet fine structure in the alkali metals, e.g., the D-lines of sodium, a spin s of $\frac{1}{2} \frac{\hbar}{2\pi}$, and a magnetic moment $\mu_s = -s 2e/2m_e c$, were ascribed to the electron as inherent properties, and a total angular momentum quantum number $j = l + s$ for a single valence electron. The resulting vector

* Sponsored by Lockheed Independent Research Program.

model of the atom is shown in Fig. 2. The magnetic interaction between the angular and spin momenta of the electron results in a coupling between l and s which then precess rapidly about their resultant j . In weak magnetic fields, j precesses slowly about H_z at the Larmor frequency times the g factor. From this vector model, we obtain

$$g = \frac{j(j+1) + s(s+1) - l(l+1)}{2j(j+1)}$$

where, in accordance with observation and the more accurate quantum mechanical deduction, j^2 , l^2 , and s^2 are replaced by $j(j+1)$, $l(l+1)$, and $s(s+1)$ respectively.

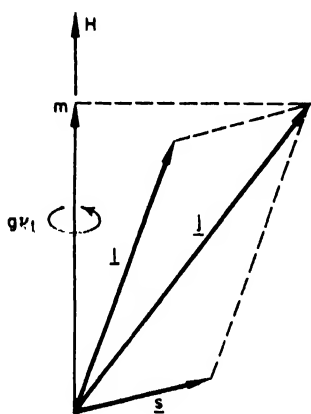


FIG. 2. Vector model of angular momenta l , s , j , and projection m of j along H for a single valence electron. Classical precession of j around H .

The magnetic quantum number m takes values j through $-j$; transitions $\Delta m = 0$ correspond to π components, $\Delta m = \pm 1$ to σ components. Selection rules $\Delta j = 0, \pm 1$ ($j = 0$ to $j = 0$ excluded) are also effective. For the transition $^2S_{1/2} \rightarrow ^2P_{3/2}$, or the D_2 -line of sodium, $l = 0$, $s = 1/2$, $j = 1/2$, $g = 2$; and $l = 1$, $s = 1/2$, $j = 3/2$, $g = 4/3$ for the lower and upper levels respectively. In terms of the normal Zeeman separation, namely, $eH/4\pi mc$, the mg values for the upper and lower levels are

$$\begin{array}{ccccccc} m & & -3/2 & -1/2 & 1/2 & 3/2 \\ mg_{\text{upper}} & \rightsquigarrow & -2 & -2/3 & 2/3 & 2 \\ mg_{\text{lower}} & \rightsquigarrow & -1 & -1/3 & 1/3 & 1 \end{array}$$

where the arrows indicate possible transitions $\Delta m = 0, \pm 1$, and give the positions of all Zeeman lines as differences between the mg terms of the two levels. These differences are $\pm 5/3$, ± 1 , $\pm(1/3)$ and give a Zeeman pattern of 6 lines for the D_2 -line of sodium (π transitions are in parentheses). Figure 3 shows the splitting and the intensities of the various transitions which depend

on the quantum numbers j and m . Anomalous Zeeman effects for other transitions and atoms are dealt with similarly. For atoms with more than one valence electron, appropriate coupling schemes between the angular and spin momenta of the electrons, such as Russell-Saunders L - S coupling, must be used in an analogous way.

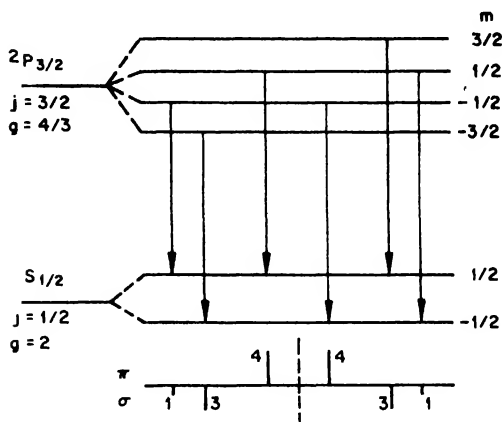


FIG. 3. Transitions involved in the anomalous Zeeman effect of the 5890 Å D_2 -line of sodium ($^2S_{1/2} \rightarrow ^2P_{3/2}$). Intensities of the π and σ components also shown.

For very strong magnetic fields, the precession frequency of j around H becomes comparable with that of l and s around j , or with the fine structure separation. This is the Paschen-Back effect region, where the fine structure coupling breaks down; l and s then precess independently around H , and the normal Zeeman pattern is observed. Paschen-Back effects have been observed in lighter elements such as Li, where the doublet separation of the 2P levels is only 0.34 cm^{-1} as compared with 17.2 cm^{-1} for the D-lines of sodium. Since 10 000 oersteds corresponds to a Larmor splitting of 0.47 cm^{-1} , very high magnetic fields are necessary with heavier atoms. In the intermediate region of magnetic fields, more elaborate quantum mechanical methods are necessary; however, a magnetic field is only strong or weak in a relative sense as compared with other internal fields in the atom.

Hyperfine structure of spectral lines involves the nuclear spin I , and a magnetic field which is weak for fine structure may be quite strong for hyperfine structure. Here the quantum number $F = I + J$ is used to determine the magnetic splitting. In optical spectra, such weak-field magnetic splittings are rarely observed because the hyperfine structure splitting is comparable with the Doppler width of the line or the instrumental resolution. Hyperfine splittings of the ground state are used in atomic beam resonance techniques to determine I , and also in atomic beam frequency standards. At higher values of magnetic field, the coupling between I and J breaks down and I and J precess independently

around H giving the Back-Goudsmit effect in hyperfine structure.

Zeeman spectra of many atoms have been observed with magnetic fields of 20 000 to 40 000 oersteds and diffraction gratings capable of resolving 1/10 of a Lorentz unit. Completely resolved spectra give the g factors and the j values of both levels involved in the transition, and the Zeeman effect has played a dominant role in the elucidation of atomic spectra.

In 1913 Stark demonstrated that every spectral line of the Balmer series of hydrogen was split into a number of components when the emitting atoms, obtained from the positive rays in a hydrogen discharge, were subjected to electric fields of 100 000 volts/cm. Lo Surdo obtained similar results using the large potential gradient in the dark space of a glow discharge. This symmetrical splitting of the hydrogen lines is due to a linear Stark effect, with a shift of the atomic levels in wave numbers given by

$$T = AF + am_m$$

Here $A = 6.45 \times 10^{-5} n(n_2 - n_1)$; F is the electric field in volts per centimeter; a is a constant; m_l and m_s are the projections of orbital and spin angular momentum on the field direction (cf. the magnetic quantum number m); n is the principal quantum number, and n_1, n_2 are other quantum numbers deduced from the Bohr quantum theory which successfully explained the effect. Certain components of the Stark pattern occur only in transverse observation and are polarized with the electric vector parallel to the field (π components). Others are polarized perpendicular to the field (σ components). For observations along the field, only the σ components appear, and these are unpolarized.

Such a linear Stark effect occurs only in hydrogen, or hydrogen-like atoms, and is due to the close proximity of levels with the same n but different angular momenta l . The quantum number m_l is analogous to the magnetic quantum number m , but with the important difference that values $\pm m_l$ now give identical splittings of energy levels. This occurs since the energy depends only on the orientation of the orbit to the electric field and not on the sense of precession as in a magnetic field. The unpolarized nature of the σ components when observed along the electric field is also due to this independence of energy on the sign of m_l .

For fields greater than 100 000 volts/cm hydrogen exhibits a quadratic Stark effect, with a shift of all spectral lines towards the red or violet.

A similar quadratic Stark effect is observed in all other atoms, where the fields are now weak compared with the separation of states of different parity, or l values. For the sodium D-lines, transitions $3s^2S_{1/2} - 3p^2P_{1/2, 3/2}$, the states arising from different electron orbits, or l values, with the same quantum number n are widely separated. These were studied by Ladenburg at fields of 160 000 volts/cm, when only a quadratic Stark effect was observed. The various components of the D-lines, which split into levels with $m_l = \pm 1/2$, and $m_j = \pm 1/2, \pm 3/2$ respectively, were shifted towards the red by a fraction of an angstrom unit. Similar results were obtained for other alkali metals. The effect increases rapidly with the total quantum number n and may be observed in absorption as well as emission. Forbidden transitions, violating the selection rule $\Delta l = \pm 1$, also occur due to a mixing of the energy levels in strong electric fields. These have been observed, particularly in helium by Foster who used electric fields up to 85 000 volts/cm.

The Stark effect is closely connected with the polarizability of the otherwise neutral atom. The various orientations of the angular momentum l then correspond to states of slightly different energies and produce the observed quadratic Stark effect. Because of its complex nature, the Stark effect, in comparison with the Zeeman effect, has not played a dominant role in the analysis of complex spectra or of atomic structure.

W. CULSHAW

References

- Born, Max, "Atomic Physics," London and Glasgow, Blackie and Son Limited, 1937.
- White, H. E., "Introduction to Atomic Spectra," New York and London, McGraw-Hill Book Co. Inc., 1934.
- Herzberg, G., "Atomic Spectra and Atomic Structure," New York, Dover Publications, 1944.
- Kuhn, H. G., "Atomic Spectra," New York, Academic Press, 1962.
- Mitchell, A. C. G., and Zemansky, M. W., "Resonance Radiation and Excited Atoms," New York, Cambridge University Press, 1934.
- Fedilov, P. P., "The Physical Basis of Polarized Emission," New York, Consultants Bureau, 1961.

Cross-references: ATOMIC SPECTRA, POLARIZED LIGHT, QUANTUM ELECTRODYNAMICS, QUANTUM THEORY, SPECTROSCOPY.

INDEX

Bold face numbers designate more important passages and small capital letters are used for titles of articles in the book.

- Abbe, Ernst, 211, 216, 429
- Abbe number, 615
- Abelson, P. H., 733
- ABERRATION, **1**, 367, **368**, 429, 480, 492
 - astronomic, 756
 - eye, 768
 - theory, **1**
- Abrikosov, A. A., 686
- Absolute pressure, **281**
- Absolute space, **619**
- Absolute temperature scale, 92
- Absorbance, 3
- Absorptance, 3
- Absorption, **13**, **18**
 - infrared, 338
 - neutron, 604
 - optical, 3
 - vibrational energy, 623
- Absorption bands, 3
- Absorption coefficient, radiation, 584
- Absorption coefficient, sound, 33
- Absorption index, 3
- Absorption peak, 460
- Absorption probability, 484
- ABSORPTION SPECTRA, **2**, 6
- Absorptive power, 3
- Accelerating frame of reference, 44
- Acceleration, 44, **179**, 404
 - angular, **632**
 - ballistics, 67
 - centripetal, **632**
 - electroluminescence, **203**
 - free fall, **133**
 - gravity, **298**
 - radial, **632**
- Accelerator
 - azimuthally varying field, 12
 - Cockcroft-Walton, 6, 9
 - cyclic, **10**
 - electrostatic, 6
 - FFAG, 12
 - fixed-field alternating-gradient, 12
- ACCELERATOR, LINEAR, **6**, 9, 469
- Accelerator, magnetic resonance, 150
- ACCELERATOR, PARTICLE, **6**, 9, **12**, 150, 459
- ACCELERATOR, VAN DE GRAAFF, **12**, 459
- Accelerometer, 333
- Acceptor, 129, 383, 641
- Acoustic attenuation, 741
- Acoustic phonon, 641
- Acoustic velocity, 740
- Acoustical branch, phonon, 507
- ACOUSTICS, **16**, 280, 534
 - Acoustics, architectural, 16, **31**
 - Acoustics, geometrical, **16**
 - Acoustics, physical, **528**
 - Acoustics, psychological, 17
- Actinide series, 733
- Action and reaction, principle of, 680
- "Action at a distance," **261**
- Activated foil, 456
- Activation analysis, neutron, 458
- Activation energy, chemical kinetics, 99
- Activation energy (diffusion), 171
- Activation factor, ion movement, insulators, **129**
- Activator, electroluminescence, **203**
- Activator ion, **383**
- Active networks, **104**
- Activity coefficient, 202, 291
- Activity, ion species, **201**
- Adams, J. C., 325
- Adatom, **690**
- Adhesion, **285**
- Adiabatic cooling (clouds), **105**
- Adiabatic demagnetization, **142**
- Adiabatic process, 91
- Adiabatic transition, 717
- Admittance, **331**
- Admolecules, 127
- Adsorbate, **18**
- Adsorbent, **18**
- Adsorption, **18**, 146
 - surface, 690
- ADSORPTION AND ABSORPTION, **17**
- Adsorption efficiency, **126**
- Adsorption isotherm, **18**
- AERODYNAMICS, **20**, 278
- Aeromechanics, 20
- Aeronomy, **296**, 345
- Aerophysics, 20
- Aerostatics, **20**, **21**
- Aether, **619**
- Affinity, electron, 737
- Afterburner, 278
- Aherns, 544
- Aiken, Howard, 124
- Air gap, magnetic circuit, 332
- Air shower, 138
- Air-breathing engine, **275**
- Airglow, **64**, 344
- Airy disc, 211
- Aizu, Kéitsiro, **255**
- Alamogordo, New Mexico, **265**

- Alchemy, 101
 Alfvén, Hannes, 393, 556, 557
 Alfvén speed, 393
 Algae, 526
 ALGOL, 125
 Alice mirror experiment, 287
 Allen, Douglas L., 751
 Alley, Charles L., 438
 Alloy, 422
 Alpha emission, 596
 Alpha particle, 469
 Alpha particle (fission), 268
 Alpha ray, 456, 467, 596
 Altair, 358
 ALTERNATING CURRENT, 22, 649
 vector diagram, 23
 Alvarez, L. W., 456
 AM, 437
 Ambipolar diffusion, 192
 Americium, 734
 Amme, Robert C., 348
 Ammeter, 193
 electrodynamic, 196
 Ampère, André Marie, 392
 Ampere (unit), 331
 determination of, 193
 Ampère's law, 205, 386, 392
 Amphoteric element, 642
 Amplification factor, photomultiplier, 520
 Amplification factor, vacuum tube, 225
 Amplifier, 626
 distributed, 497
 Amplitude, 453, 645, 760, 775
 wave, 775
 Amplitude modulation, 437
 Analogue computers, 123
 Analysis, activation, 458
 Analysis, tracer, 594
 Analytical chemistry, 71, 102
 Analytical engine, 124
 Analyzer, differential, 124
 Analyzer, polarized light, 512
 Anaximenes of Miletus, 233
 Anderson, C. D., 29, 210, 549
 Anderson, David L., 210
 Anderson, Ernest R., 666
 Anderson, J. S., 169
 Andrews, C. L., 178
 Ångström, A. J., 741
 Ångström unit, 59, 741
 Angular acceleration, 632
 Angular dispersion, 495
 Angular impulse, 327
 Angular momentum, 45, 237, 327, 455, 480, 531, 632, 753
 conservation, 130, 327
 electron, 387
 molecular, 70
 orbital, 501
 particle, 684
 spin, 256
 Angular velocity, 135, 753
 Anion, 200, 201, 346
 Anisotropic crystal, 616
 Anisotropy, magnetic, 28, 257
 Anisotropy, magnetocrystalline, 390
 Annealing, 170
 irradiation effects, 351
 Annihilation, positron-electron, 550
 Annihilation radiation, 549
 Anode, 201, 346
 vacuum tube, 222
 Anode fall, 192
 Anomalous skin effect, 253
 Ansatz, 546
 Anschütz-Kämpfe, Hermann, 301
 Antarctica, 344
 ANTENNA, 25, 436, 576
 Antiferroelectric, 255
 Antiferromagnetic rotation, 249
 ANTIFERROMAGNETISM, 28, 256, 390, 462
 Antikaon, 232
 Antineutrino, 455, 469, 596
 ANTIPARTICLE, 29, 131, 263, 522, 549
 Antireflection coating, 614, 723
 Anti-Stokes' emission, 367, 381, 599
 Antivibration mounting, 416
 Anvil, Bridgman, 552
 Aperture, numerical, 430, 481
 Aperture, relative, 481
 Aperture stop, 368
 Aperture synthesis, 591
 Aphelion, 357
 Apoapsis, 37
 Apocenter, 48
 Apogee, 359
 Apostilb, 379
 Appleton-Hartree equation, 555
 Aqueous humor, 766
 Aquinas, St. Thomas, 324
 Arago, François, 757
 Arc, 193, 371
 Archimedes, 679
 Archimedes' principle, 21, 157, 282
 ARCHITECTURAL ACOUSTICS, 16, 31
 Archytas of Taras (or Tarentum), 679
 Arecibo, radio telescope, 591
 Aristotle, 324
 Armature, 183, 451
 Armstrong, H. L., 676
 Array antenna, 576
 Arrhenius, Svante August, 78, 98, 200, 440, 530
 Arrhenius relation, 171
 Arrhenius' theories of reaction rates, 530
 Artificial element, 234
 Artificial horizon, 301
 Artificial radioactivity, 596
 Ashby, Ross, 149
 Ashby, W. R., 249
 Aspheric surface, 367
 Assembly, thermodynamics, 719
 Astatic design, 417
 Astatic magnetometer, 395
 Astigmatism (lens), 492
 Aston, Francis William, (isotope), 352, 404, 476
 ASTRODYNAMICS, 36
 ASTROMETRY, 41
 ASTRONAUTICS, PHYSICS OF, 44
 Astronomical telescope, 480
 Astronomical time, 725
 Astronomy, 41, 652
 geodetic, 294
 radio, 591
 ASTROPHYSICS, 48
 Atactic polymer, 547
 Atkinson, L., 451
 Atmosphere, 344, 535
 absorption in, 49
 chemistry, 348
 earth, 348, 666
 fallout, 247
 planets, 71, 535, 666
 pressure, 281
 windows, 338
 Atom, 51, 57, 100, 228, 236, 362, 386, 444, 476, 501, 530
 Bohr, 236
 displaced, 350

- protons in, 559
- size, 504
- Atomic absorption coefficient, 5
- ATOMIC AND MOLECULAR BEAMS, 51
- Atomic bomb, 265
- ATOMIC CLOCK, 53, 132, 485
- ATOMIC ENERGY, 55, 264, 470
- Atomic Energy Commission, 265
- Atomic gyroscope, 485
- Atomic mass scale, 439
- Atomic mass unit, unified, 133
- Atomic number, 58, 62, 234, 458, 502
- ATOMIC PHYSICS, 57
- Atomic radius, 504
- Atomic second, 54
- ATOMIC SPECTRA, 59, 671
- Atomic structure, 57, 502
- Atomic weight (transuranium elements), 733
- Attenuation, acoustic, 741
- Attenuation calculation, shielding, 604
- Attenuation, neutron, 605
- Attitude (guidance), 336
- Auditoriums, acoustics of, 34
- Auditory behavior, 309
- Auditory cortex, 310
- Auditory pathways, 309
- Auger, Pierre, 63
- AUGER EFFECT, 62
- Auger electrons, 63
- Auger spectroscopy, 63
- Auger yield, 63
- Augustine, St., Bishop of Hippo, 324
- Auricle, 309
- Aurora, 64, 138, 296, 344, 349, 394, 667
- radio, 65
- AURORA AND AIRGLOW, 64
- Aurora borealis, 64, 394
- Auroral zone, 64
- Autodin, 708
- Autoelectronic emission, 258
- Automatic volume control, 226, 249
- Automation, 249
- Avalanche emission, 204
- Avalanche transistor, 562
- Averaging, 415
- AVF cyclotron, 12
- Avogadro, Amedeo, 101, 325, 438, 441
- Avogadro constant, 133
- Avogadro's hypothesis, 102, 441
- Avogadro's number, 86, 201, 438, 439, 459
- Avogadro's principle, 439, 445
- Axford, W. I., 668
- Axial aberration, 368
- Axial point group, 147
- Axis, optic, 496
- Axle, 647
- Axon, 767
- Azimuthally varying field accelerators, 12
- Babbage, Charles, 124
- Back, E., 220, 791
- Backscattering, 635
- Bacon, G. E., 463
- Bacon, Roger, 429
- Bacteria, 523
- Bahr, Eva von, 68
- Balescu, Radu C., 683
- Ballistic pendulum, 67
- BALLISTICS, 36, 67
 - Coriolis effect, 135
 - exterior, 67
 - interior, 67
 - terminal, 67
- Balmer, J. J., 59
- Balmer series, 60
- Band
 - conduction, 510
 - energy, 154
 - valence, 510
- Band, William, 126, 409
- BAND SPECTROSCOPY, 68, 672
- Band structure, 304, 640, 663
 - effect on reflectivity, 611
- Bandwidth, oscilloscope, 498
- Barbrow, L. E., 519
- Bardeen, J., 686, 728
- Barium titanate, 739
- Barkhausen effect, 390
- Barkla, Charles Glover, 63, 120, 784, 785
- Barn, 142, 267, 457
- Barometer, mercury, 282
- Barometric pressure, 282
- Barometric-height relation, 283
- Bartlett, Neil, 102
- Barycentric coordinate system, 110
- Baryon, 229, 231
- Baryon number, conservation of, 131, 229
- Base (transistor), 728
- Basilar membrane, 310
- Batholinus, Erasmus, 496
- BATTERY, 72, 200, 328
 - storage, 72
- Bay, Z., 591
- Bayard-Alpert gauge, 748
- BCS theory, 686
- Beam, atomic, 51
- Beam, molecular, 51, 54
- Beam-power tube, 225
- Beams, J. W., 95
- Beccari, E., 408
- Beck, Clifford K., 474
- Becker, Joseph J., (magnetism), 391
- Becquerel, Antoine Henri, 78, 101, 306, 325, 468, 595
- Beer, A., 3 (footnote)
- Beer, Albert C., 306
- Beer's law, 3 (footnote)
- Bel, 17, 453
- Bell, Alexander Graham, 17
- Bell, D. A., 150
- Bellman, P., 408
- "Belt," General Electric, 552
- Rénard cell, 394
- Bennet, W. R., 366
- Bennett, H. E., 614
- Benoit, J. R., 343
- Benumof, Reuben, 25
- Beran, Mark J., 107, 109
- Beranek, Leo L., 35
- Bergstrand, Erik, 758, 759
- Bergstrom, Auger effect constants, 64
- Berkelium, 735
- Berman, Arthur I., 48
- Bernoulli, Daniel, (kinetic theory), 361
- Bernoulli's equation, 20, 280
- Beryllium, 456
- Besançon, Robert M., 535, 711
- Bessel function, 650
- Beta decay, 268, 455, 477, 596, 778
- Beta ray, 467, 596
 - spectroscopy, 671
- Beta-particle, 264
- BETATRON, 9, 75, 469, 704
- Betatron oscillations, 75
- Betelgeuse, 358
- Bethe, Hans, (stopping power), 637
- Bevel gear, 649

- Beyer, George L., 444
 Bimetallic strip, 711
 Bimirror, 340
 Binary stars, 43
 Binder, Raymond C., 283
 Binding energy, 477
 Binding energy (polaron), 546
 Biochemistry, 102
 Biological processes, 78
 Biology, mathematical, 506
 Biomathematics, 406
 Biomedical engineering, 420
 BIONICS, 76
 BIOPHYSICS, 78, 420
 Biophysics, mathematical, 406
 Biophysics, molecular, 79
 Biprism, 340
 Birefringence, electric, 174, 512, 544, 616
 Birss, R. R., 398
 Bjerknes, Vilhelm, 295
 Black body, 3, 286, 337, 516, 563, 587
 radiation, 585
 Black lighting, 744
 Blackman, 155
 Blankenbecker, R., 112
 Blasius of Parma, 679
 Blewett, John P., 8
 Blewett, M. Hildred, 12
 Blizard, Everitt P., 605
 Bloch, F., 456
 Bloch-Grüneisen formula, 129
 Blocking oscillator, 561
 Bloembergen, N., 169, 372
 Boast, Warren B., 551
 Bode, H. W., 149
 Bode plot, 644
 Body, fluid, 178
 Body, rigid, 178
 Boersch, H., 215
 Bohm, Henry V., 254
 Bohr, Niels, 101, 208, 236, 250, 387, 400, 501, 530, 771
 fission, 264
 nucleus, 478
 quantization, 58
 quantum theory, 713
 stopping power, 637
 Bohr atom, 236
 Bohr frequency, 773
 Bohr magneton, 134, 175, 219, 386, 389, 501
 Bohr model, 559
 Bohr radius, 134
 Bohr-Sommerfeld theory (spectra), 60
 Boldwood, 352
 Bolometer, 338, 590
 Boltzmann, Ludwig, 80, 92, 250, 319, 325, 362, 532, 762
 distribution function, 245
 Boltzmann constant, 80, 83, 134, 250, 289, 319, 501
 Boltzmann distribution, 240, 245, 388
 BOLTZMANN DISTRIBUTION LAW, 80
 Boltzmann equation, 363
 Boltzmann factor, 715
 Boltzmann *H*-theorem, 242, 363
 Boltzmann hypothesis, 242
 Boltzmann statistics, 401, 533
 Boltzmann superposition principle, 762
 Bomb, atomic, 265
 Bomb, hydrogen, 265
 Bomb calorimeter, 89
 Bond, chemical, 71, 81, 440, 444, 504
 coordinate-covalent, 444
 dipole moment, 83
 dissociation, 82
 electrovalent, 444
 energies, 82
 homopolar, 444
 length, 82, 445
 moment, 174
 semipolar, 444
 surface, 690
 Bondi, H., 139, 140
 Bondi and Gold theory, 140
 Bonding, chemical. *See* Bond
 Bonding (friction), 283
 Booker, 557
 Born, Max, 210, 773
 Born approximation, 85
 Born-Oppenheimer approximation, 446
 Bose, Sir Jagadis Chandra, 251, 815
 Bose-Einstein distribution, 83
 ROSE-EINSTEIN STATISTICS AND BOSONS, 83, 251, 263, 522, 682, 688
 Bosons, 83, 251, 682
 Boston Symphony Hall, 35
 Bottle, magnetic, 56
 Bouguer, Pierre, 3 (footnote)
 Bouguer and Lambert, law of, 3
 Boulengé chronograph, 67
 Boundary layer, 21
 Box, one-dimensional, 772
 Boyle, Robert, 101, 233, 289, 361
 Boyle's law, 289
 Boynton, Robert M., 768
 Braddick, H. J. J., 418
 Bradley, C. S., 451
 Bradley, James, 756
 Bradley, John N., 646
 Bragg, W. H., and W. L., 5, 752, 783, 785
 reflection planes, 252
 Bragg's law, 783
 Brahe, Tycho, 357
 Branching ratio, 354, 597
 Brandstatter, Julius J., 559
 Brattain, W. H., 728
 Bravais lattice, 148
 Brayton cycle, 277
 Breakdown, potential, 191
 Breakdown, spark, 191
 Breeder reactor, 734
 BREMSSTRAHLUNG, 84, 286, 541, 582, 584, 741
 Breneman, John W., 649
 Brett, J., 450
 Brewster, Sir David, 496, 544
 Brewster's angle, 544
 Brewster's law, 496
 Bridge, electrical, 193, 194
 Bridge, impedance, 194
 Bridge, Kelvin, 194
 Bridge, Wheatstone, 194
 Bridgman, P. W., 552
 Bridgman anvil, 552
 Bridgman technique, 146, 639
 Briggs, G. E., 500
 Brightness, 115, 379
 electroluminescence, 204
 Brightness temperature, 563
 Brillouin function, 501
 Brillouin zone, 252, 553
 British thermal unit, 312
 Brittle behavior, 423
 Brown, F. C., 546
 Brown, Robert, 86
 BROWNIAN MOTION, 86, 94, 407
 Brüche, E., 212
 Brueckner, K. A., 400
 Brueckner theory, 399, 400

- Brunauer-Emmet-Teller method, 19
 Brush, Stephen G., 364
 Btu, 311
 Buchdahl, H. A., 720
 Bueche, F., 548
 Buildup factor, 604
 Bunson ice calorimeter, 88
 Buoyant force, 754
 Burgers vector, 127 (footnote)
 Burgh, Donald A., 433
 Burhop, E. H., 64
 Burkhard, Donald G., 436
 Burning rate, 67
 Burst, weapon, underground, 247
 Bush, Vannevar, 124
 Butsch, Leonard M., Jr., 77
 Button cell, 74

C region, 348
C. S. C. O., 30
 Caianiello, E. R., 264
 Calcite, 544
 Caldwell, 124
 Californium, 735
 Callen, Earl, 398
 Caloric, 311
 Calorie, 312
 international steam table, 133
 thermochemical, 133
 Calorimeter, 87, 89
 CALORIMETRY, 87, 292
 Calutron, 352
 Camera, 480, 514
 Camera, objective lens, 369
 Campani, 329
 Campbell, N. R., 635
 Candela, 379, 516
 Cannizzaro, Stanislav, 441
 Cannon, E. W., 125
 Canonical distribution, 682
 Canonical ensemble, 241
 Canonical variable, 565
 CAPACITANCE, 23, 89, 159
 measurement of, 193
 self-, 331
 Capacitor, 89, 193, 622
 Capillarity, 692
 Capture, electron, 597
 Capture, gamma ray, 597, 604
 Carathéodory, Constantin, 717
 principle of, 717
 Carbon cycle, solar, 651
 Carburetor, 280
 Cardiovascular phenomena, 407
 Carlisle, Anthony, 200
 Carnot, Sadi, 92, 325
 Carnot cycle, 711
 CARNOT CYCLES AND CARNOT ENGINES, 91
 Carnot engine, 91, 711
 Carnot refrigerator, 91
 Carrier, hot, 204
 Carrier, minority, 610
 Carrier, photoconductivity, 510
 Carrier rotation, 248
 Carrier system, telegraphy, 707
 Carrier wave, 437
 Cartesian axes, 410
 Carver, Thomas R., 485
 Cascade, 584
 Cascade process, liquefaction, 374
 Casey, E. J., 80
 Casting, 421
 Catalogues, star, 41
 Catalyst, 98, 99, 472

 Cathode, 201, 346
 oxide-coated, 222
 vacuum tube, 222
 Cathode fall, 192
 Cathode ray, 208, 380
 Cathode-ray oscilloscope, 496
 Cathode-ray tube, 217, 496
 Cathode sputtering, 721
 Cathodoluminescence, 380
 Cation, 200, 201, 346
 Cat's whisker, 172
 Cauchy, A. L., (optics), 3
 Cavendish, Henry, 323
 CAVITATION, 93, 528
 Celestial mechanics, 36, 41
 Cell
 button, 74
 concentration, 202
 cylindrical, 74
 electric, 200
 electrochemical, 72
 fuel, 75
 galvanic, 72
 Leclanché, 72
 primary, 74
 receptor, 77
 rechargeable, 73
 rectangular, 74
 secondary, 74
 standard, 74
 unit, 460
 voltaic, 72
 "Weston," 74
 Cell division, 407
 Cell structure, 378
 Celsius, Anders, 710
 Celsius scale, 710
 Center of gravity, 680
 Center of mass, 680
 Centered lattices, 148
 Centigrade absolute scale, 710
 Centigrade scale, 710
 Central force field, 60
 Central force motion, 44
 Centrifugal force, 93, 317, 633
 Centrifugal stretching, 434
 CENTRIFUGE, 93, 354
 Centripetal acceleration, 44, 632, 753
 Centripetal force, 633
 Centroid, 680
 Ceramic transducers, 199
 Čerenkov, P. A., 96, 97
 Čerenkov counter, 467
 ČERENKOV RADIATION, 96, 380
 Cerium earths, 601
 "Cerium" metal, 602
 Cesium clock, 51, 53, 132
 Chadwick, Sir James, 456, 474
 Chain reaction, 55, 56, 264, 266, 470, 471
 polymers, 581
 Chako, Nicholas, 2, 570
 Chalmers, 355
 Chamber, rocket, 272
 Chancourtois, 502
 Chandrasekhar, B. S., 687
 Change of state, 675
 Chapman, Sidney, 295, 362
 Characteristic curve (semiconductor), 173
 Characteristic x-ray, 238
 Charge
 dissipation, 678
 electric, 550, 678
 electrical, 550, 678
 elementary, 133

- Charge (Cont.)**
 neutralization, 679
 space, 192
 surface, 692
 Charge carrier, 203, 303
 Charge conjugation, 131
 Charge density, 550
 Charge exchange, 347
 Charge exchange (fusion), 286
 Charge independence, 231
 Charge to mass ratio for electron, 134, 208
 Charge transfer, 346
 Charge separation mechanism, 527
 Charles, Jacques, 289
 Chemical analysis, 71
 Chemical bond, 71, 81
 Chemical detector, 467
 Chemical diffusion, 170
 Chemical elements, 233
 Chemical equilibrium, 241
 Chemical equivalent weight, 201
 CHEMICAL KINETICS, 98
 CHEMICAL PHYSICS, 100
 Chemical potential, 241, 718
 Chemiluminescence, 380
 Chemisorption, 19
 CHEMISTRY, 101
 Chemistry, physical, 530
 Chew, G. F., 111, 112
 Chlorophyll, 525
 Choke coil, 331
 Chopper amplifier, 417
 Choppin, 736
 Christie, Dan E., 755
 Christofilos, Nicholas C., 704
 Chromatic aberration, 367
 Chromatography, 101, 533
 Chromosphere, 660
 Chronograph, Boulengé, 67
 CIE, 113, 516
 C-invariance, 31
 Circuit, integrated, 144
 Circuit, nonlinear, 104
 Circuit, tuning, 24
 CIRCUITRY, 102
 Circular motion, 631
 Circular polarization, 543
 Circulator, 254
 Cladding, 473
 Clapeyron, E., 92, 377, 749
 Clapeyron relation, 241, 377
 Clark, George L., 785
 Claude, George, 376
 Claude cycle, 376
 Claude process, 375
 Clausius, Rudolf Julius Emanuel, 92, 174, 239, 362, 541, 749
 Clausius-Clapeyron equation, 749
 Clausius and Mosotti equation, 246, 541
 Cleavage plane, 690
 Clinical thermometer, 711
 Clipping amplifier, 561
 Clock, 724
 atomic, 53, 132
 cesium, 53, 132
 Close, Kenneth J., 749
 Closed sub-shell, 61
 Cloud, ionic, 201
 Cloud chamber, 465, 467, 549
 CLOUD PHYSICS, 105, 425
 Cloud seeding, 106, 425
 Cluster expansion, 682
 Coactivator, 383
 Coating, antireflection, 723
 Coating, optical, 722
 Coaxial line, 776
 Coblentz, W. W., 71
 COBOL, 125
 Cochlin, Ira, 302
 Cochlea, 310
 Cockcroft, J. D., 479
 Cockcroft-Walton accelerator, 6, 9
 Coefficient of expansion, 244
 Coercive force, 257, 390
 Cohen, E. Richard, 133, 135
 Cohen, V. W., 456
 COHERENCE, 97, 106, 365
 superconductivity, 687
 Coherence length, 686
 Coherent light, 371
 Coherent radiation, 365
 Coil, inductance, 331
 Cold emission, 258
 Cold working, 422
 Cole, Kenneth, 407
 Collective description, 399
 Collective model, nucleus, 475
 Collective motion theory, 399
 Collector (transistor), 728
 Colligative methods, molecular weight, 441
 Collins, S. C., 142
 Collision, 730
 elastic, 190
 focusing, 350
 neutrons, 473
 superelastic, 347
 Collision cross section, 287, 484
 Collision-free mirror, plasma, 287
 COLLISIONS OF PARTICLES, 109, 347, 539, 635
 Colloid, electrical propulsion, 188
 COLOR, 113, 370
 tone, 454
 COLOR CENTERS, 116
 in semiconductor, 118
 Color photography, 515
 Color scales, 115
 Color temperature, 587
 Color vision, 78, 768
 Colorants, 115
 Color-matching functions, 113
 Column, positive, 192
 Column vector, 409
 Coma, 368
 Combined inversion, 229, 779
 Communication, telemetry, 709
 Commutation relation, 566
 Commutator, bar, 610
 Commutator, generator, 183
 Commutator, matrix mechanics, 772
 Comparator technique (activation analysis), 459
 Compatible observables, 30
 Compensation point, 254
 Completely defined state, 80
 Complex, molecular, 445
 Complex compliance, 762
 Complex modulus, 762
 Component, vector, 753
 Components (thermodynamic), 506
 Composition of forces, 680
 Compound, 233, 507
 molecular, 445
 Compound nucleus, 266, 470
 COMPRESSIBILITY, GAS, 118
 Compressibility factor, 118, 291
 Compression, 281, 292, 314, 377, 528, 553
 heat of, 119
 Compressional wave, 775
 Compton, Arthur Holly, 120, 307, 785

- Compton, W. Dale, **118**
 COMPTON EFFECT, **120**, 321, 468, 785
 proton, **122**
 proton and deuteron, **122**
 Compton process, 584
 Compton scattering, 380, 465, 583
 gamma, 604
 Compton shift, **121**
 Compton wavelength, electron, **120**, **134**
 Compton wavelength, proton, **134**
 COMPUTER, **123**
 superconducting, 687
 transistorized, 729
 Concave lens, 492
 Concentration cell, **72**, **202**
 Concentration energy, solar energy, 655
 Concentration polarization, electrolysis, **202**
 Concentration ratio, solar energy, 655
 Concert hall, acoustics of, **35**
 CONDENSATION, 105, 119, **126**, 425
 Condensation coefficient, **126**
 Condenser, lens, **369**
 Condon, E. U., 596
 Conductance, electrical, 88, **127**
 Conductance, solutions, 201
 Conductance, specific (dielectrics), **160**
 Conduction, heat, **315**
 Conduction band, 82, **129**, 336, 510, 640, 649, 662
 CONDUCTIVITY
 ELECTRICAL, **127**, 510, 662, 679
 electrolyte, **346**
 equivalent, **346**
 semiconductor, 641
 thermal, **316**, 508, 662
 Conductivity tensor, 304
 Conductor, electrolytic, 200
 Conductor, electronic, 200
 Conductor, irradiation effects, **351**
 Cone (eye), 767
 Congruent melting point, 507
 Conjugate foci, **488**
 Conjugation, 522
 Conjugation, particle, **229**
 Conjugation operator *C*, **29**
 Conservation
 of baryon number, **131**
 of electric charge, **130**, **229**
 of energy, 419, 781
 of isotopic spin, 231
 of momentum, 270, **326**
 of strangeness, **131**
 Conservation law, **29**, 130, 228, 229
 angular momentum, **130**
 energy, **130**
 momentum, **130**
 multiplicative, 229
 CONSERVATION LAWS AND SYMMETRY, **130**
 Consolute point, 507
 Constancy of interfacial angles, law of, **147**
 Constant, radiation, first, **134**
 CONSTANTS, FUNDAMENTAL, **132**
 Constituents of matter, **231**
 Constitutive coordinates, **718**
 Constructive interference, **493**
 Contact charging, **677**
 Continuity equation, **20**, 528, 754
 Continuum dynamics, **279**
 Contraction, lanthanide, **601**
 Control, automatic volume, 249
 Control (Cybernetics), **148**
 Control, thrust vector, **275**
 Control grid, **225**
 Control systems, **149**
 Controlled rectifier, 562, 606
 Controller, 642
 Convection, 88, 315, **317**
 Conversation (feedback), **249**
 Converter, 305
 Convex lens, 492
 Cook, C. Sharp, **247**
 Cooke, Sir William Fothergill, 450
 Cooper, L. N., 686
 Coordinate, thermodynamic, 716
 Coordinate-covalent bond, 444
 Coordination number, 378
 Cope, Freeman, 406
 Copernicus, Nicolaus, 324, 357, 429
 Copolymer, 548
 Corbin, John C., Jr., **650**
 Corbino, O. M., **306**
 Corbino effect, **306**
 Corbino magnetoresistance, **306**
 Coriolis, G., 336
 Coriolis acceleration, **135**
 Coriolis component of acceleration, 179
 CORIOLIS EFFECT, **135**
 Coriolis force, **135**, 317, 336, 633
 Corngold, N. R., 456
 Corona, 323, **660**, 766
 Corrosion, fatigue, 423
 Corrosion, metals, 423
 Corti, organ of, 310
 Cosine law, Lambert, **519**
 COSMIC RAY, **137**, 344, 578, 668
 Cosmological constant, 139
 Cosmological principle, 139
 COSMOLOGY, 49, **139**, 300, 357, 409
 Coster-Kronig effect, **63**
 Cotton, H., **493**
 Cotton-Mouton effects, 175
 Cottony, H. V., **28**
 Couette flow, **279**
 Coulomb, Charles Augustin de, 550
 Coulomb energy, 81
 Coulomb equivalent, **201**
 Coulomb field, 60
 Coulomb force, 263
 Coulomb interaction, 582
 Coulomb potential, 546
 Coulomb's law, **197**, 228, 261
 Counter, proportional, 465, **466**
 Counter, scintillation, 459, 522
 Couper, Archibald, 101
 Couple, 680
 Coupling, electromechanical, **199**
 Coupling, spin-orbit, **61**
 Courant, E. D., 704
 Covalent bond, **81**
 Covariance, 573
 Coveyou, R. R., 307
 Cowan, C. L., **778**
 Crab Nebula, **138**, 544
 Cranberg, L., 603
 Crane, H. R., 220
 Crawford, Franzo H., **313**
 Creation operator, **30**
 Creep, **423**
 Creep compliance, 762
 Critical current, 686
 Critical field, superconductivity, **685**
 CRITICAL MASS, 55, **140**, 265, 266, 478, 731
 Critical point, 507, **675**
 Critical pressure, **675**
 Critical size, chain reaction, 141
 Critical temperature, **675**, **749**
 Criticality factor, **266**
 Crookes, Sir William, 208
 Crookes cathode-ray tube, 306, 784

- Cros, Charles, 625
 Cross section, 141, 457, 459, 473, 475, 730
 atomic, 483
 chemical kinetics, 98
 collision, 287, 484
 scattering, 109
 CROSS SECTION AND STOPPING POWER, 141
 Cross-relaxation, 402
 Cryogenic pumping, 748
 CRYOGENICS, 142
 Cryohydric point, 507
 Cryotron, 144
 Crystal, 146, 460
 equilibrium shape, 144, 690
 ferroelectric, 255
 growth, 144
 host (color centers), 117
 isotropic, 616
 lattice, 147, 783
 optical rotation, 432
 structure, 147, 661
 x-ray diffraction by, 783
 Crystal defect, diffusion of, 171
 Crystal symmetry, 147
 Crystalline lens, 766
 Crystallinity, 548
 CRYSTALLIZATION, 144, 769
 CRYSTALLOGRAPHY, 146, 661
 Crystallometry, 147
 Culshaw, W., 792
 Curie, Irene, 354, 468
 Curie, Marie Sklodowska, 306
 Curie, Pierre, 389, 501
 Curie (unit), 597
 Curie constant, 501
 Curie law, 390, 501
 Curie point, 254
 Curie temperature, 256, 389, 501
 Curie-Weiss law, 246, 501
 Curie-Weiss temperature, 501
 Curium, 57, 734
 Curl V, 754
 Current
 alternating, 22
 effective, 22
 electric, 102, 127, 183, 197, 327, 330, 649
 electric, direction of, 200
 magnetic, 328
 magnetic field due to, 198
 measurement of, 193
 particles, 303
 positive, 191
 Current density, 385, 551
 Current transformers, 196
 Curtis, L. F., 458
 Curved space, 299
 CYBERNETICS, 148, 249, 409
 Cyclic accelerators, 10
 CYCLOTRON, 10, 150, 704
 AVF, 12
 sector-focused, 12
 Cyclotron frequency, 153
 Cyclotron mass, electron, 156
 Cyclotron radius, 213
 CYCLOTRON RESONANCE, 153, 158, 253, 399
 Cyclotron resonance condition, 150
 Cylindrical cell, 74
 Czerlinsky, Ernst R., 397
 Czocharlski technique, 146, 639
 D. C. X. mirror experiment, 287
 D region, 348
 da Vinci, Leonardo, 679
 d'Alembert principle, 240
 Dallas, A. E. M. M., 63
 Dalton, John, 440, 750
 Dalton's law, 750
 Damping, 761
 Danielli, J. F., 406, 407
 Danziger, L., 408
 Dark current, 418
 Dark field microscopy, 432
 Dark ground illumination, 483
 Dark noise, photomultiplier, 522
 Dark space, Faraday, 192
 Dauvillier, A., 742
 Davis, R. H., 560
 Davis, Sumner P., 361
 Davisson, Clinton Joseph, 209, 210, 531, 771
 diffraction, 163
 Davy, Sir Humphry, 200
 Davydov splitting, 243
 Day, Jack E., 498
 Dayglow, 65
 Dean, L. Wallace, III, 530
 de Broglie, Louis Victor, 101, 209, 210, 321, 531, 571, 771
 de Broglie hypothesis, 216, 634
 de Broglie wavelength, 211, 259, 531
 Debrunner, Peter B., 450
 Debye, Peter J., 142, 159, 201, 348, 540
 dielectrics, 160
 diffraction, 164
 polarization, 174
 Debye heat capacity, 166, 314
 Debye length, 540
 Debye shielding distance, 540
 Debye temperature, 128
 Debye-Hückel radius, 348
 Debye-Hückel theory, 201, 440, 533
 Decay mode, 228
 Decay rate, 597
 Decay scheme (radioactivity), 597
 Deceleration radiation, 204
 Déchêne, 204
 Decibel, 17, 310, 453, 665
 Dee, 150
 Defares, J., 408
 Defect, lattice (semiconductors), 128
 Defect, mass, 477
 Defects, mirrors and lenses, 492
 Deflection, electrostatic, 218
 Deflection plate, 497
 Deflector, cyclotron, 151
 Deformation, 762
 Deformation and flow, 630
 Degeneracy, 60
 Degenerate energy level, 80
 Degree, Kelvin, 133
 Degree of freedom, 80, 119, 387, 506
 de Haas, W. J., 144, 155
 DE HAAS-VAN ALPHEN EFFECT, 144, 155, 253
 de Haas-van Alphen experiments, 154
 Delayed neutron, 266, 475, 478, 603
 Delocalization energy, 82
 Delta function, 409, 565
 Delta ray, 583
 Delvaille, John P., 138
 Demagnetization, adiabatic, 142
 Dember, H., 511
 Dember voltage, 511
 Democritus, 324, 361
 Demodulation, telemetry, 709
 Dempster, A. J., 352
 Density, 279, 528
 DENSITY AND SPECIFIC GRAVITY, 156
 Density gradient tube, 157
 Density operator, 681

- Depth of field, 431
- Descartes, René, 261, 614
- Desroziers, 450
- Destriau, G., 203
- Destriau effect, 203
- Destruction operator, 30
- Destructive interference, 493
- Detector
 - infrared, 338
 - quantum, 590
 - radiation, 464
 - semiconductor, radiation, 467
- Deuterium, 285, 461, 559
- Deuterium-deuterium reactions, 56
- Deuterium-tritium reaction, 56
- Deuteron, 457, 559, 733
 - acceleration of, 704
- Deutsch, M., 122
- Developer, photographic, 514
- Deviation, minimum, 487
- Deviation, optics, 485
- Device, thermonuclear, 265
- Devices, microwave, 249
- Dewar, Sir James, 375
- Dexter, David L., 664
- Diagonal matrix, 409
- Diamagnetic resonance, 158
- DIAMAGNETISM, 158, 389
- Diamond, synthesis of, 552
- Diatomic molecule, 68, 69
- Dichroic polarizer, 544
- Dichromatic vision, 113
- Dicke, R. H., 140, 299
- Didymium, 4, 602
- Diebold, Edward J., 610
- Dielectric, 90, 159
- Dielectric coefficient, 128
- Dielectric constant, 90, 159, 201
 - optical, 159
 - polymers, 548
- Dielectric film, 723
- Dielectric loss, 159, 542
- Dielectric medium, 242
- Dielectric polarization, 205
- Dielectric relaxation, 159
- DIELECTRIC THEORY, 159
- Dienes, G. J., 351
- Differential analyzer, 124
- Differential pulley, 647
- Diffraction, 341, 370, 494, 771
 - electromagnetic waves, 207
 - electron, 210
 - neutron, 460
 - x-ray, 442
- Diffraction grating, 160, 494
- DIFFRACTION BY MATTER AND DIFFRACTION GRATINGS, 160
- Diffuse reflectance, 613
- Diffuse series, 61
- Diffusion, 146, 168
 - ambipolar, 192
 - ions, 192
 - IN LIQUIDS, 168
 - molecular, 749
 - solid state, 595
 - IN SOLIDS, 170
 - thermal, 362
 - isotope, 354
- Diffusion coefficient, 169, 170
- Diffusion drag force, 407
- Diffusion pump, 747
- Diffusivity, thermal, 316
- Digital computers, 124
- Dinsdale, A., 765
- Diode, 606
 - junction, 172
 - point contact, 172
 - (SEMICONDUCTOR), 172
 - snap, 562
 - tunnel, 562
 - vacuum tube, 222, 224
- Diode-triode, 223
- Diopter, 493
- Dipole, 527
 - acoustics, 529
 - antenna, 26
- Dipole interaction, 435, 485
- DIPOLE MOMENT, 159, 173, 257, 378, 435, 541
- Dipole moment, magnetic, 174, 387
- Dipole-dipole force, 257
- Dirac, Paul Adrian Maurice, 58, 81, 120, 140, 209, 219, 250, 262, 409, 568, 771 (footnote)
- Dirac delta function, 207
- Dirac energy states, 549
- Dirac equation, 219, 262, 568
- Dirac field, 565
- Dirac matrices, 568
- Dirac relativistic anti-electron theory, 29
- Dirac theory of electron, 120
- Direct-current generator, 183
- Direct-current instruments, 195
- Directional coupler, waveguide, 778
- Disc, Airy, 211
- Discharges
 - arc, 193
 - in gases, electrical, 189
 - glow, 192
 - high current, 193
- Disintegration, nuclear, 457
- Dislocation, 260, 661, 741
- Dislocation lines, 422
- Dispersion, 494, 625
 - angular, 495
 - optical, 542, 615
- Dispersion curve, 495
- normal, 3
- Dispersion hardening, 422
- Dispersion relation, 264, 555
- Dispersion theory, 112
- Dispersive power, 495, 615
- Displacement (vibration), 761
- Displacement current, 385, 554
- Displacement law, Wien, 586
- Displacement vector, 181
- Dissipation factor, 159
- Dissociation, 346
 - of molecules, 584
- Distances, interatomic, 71
- Distortion, amplifier, 498
- Distortion, lens, 368
- Distributed amplifier, 497
- Distribution function, 681
- Divergence, 754
- Divider, voltage, 194
- Dolland, John, 429
- Domain, magnetic, 256, 390
- Donor, 129, 383, 641
- Doping, 172
- Doppler, Christian, 176
- DOPPLER EFFECT, 176, 483, 541, 576, 776
- Doppler shift, 43, 50, 54, 447
- Doremus, Robert H., 146
- Dosimetry, 580
- Double refraction, 360, 496, 616
- Double stars, 43
- Doublet structure, 61
- Downey, Glenn L., 180
- Drag (aerodynamics), 20, 271

- Drebbel, Cornelis, 429
 Dresselhaus, G., 154
 Drickamer, H. G., 554
 Drift, electron, 190
 Drift tubes, 7, 10
 Drift velocity, charge carriers, 128
 Drude, Paul Karl Ludwig, 422
 electric current, 714
 Duane, William, 785
 Dulong and Petit rule, 315
 Dumas, Jean Baptiste André, 441
 DuMond, J. W. M., 133
 Dunlap, W. Crawford, 642
 Dunover, 51
 Dushman, S., 224, 715
 Dwight, C. Harrison, 6, 496
 Dwight, H. B., 650
 Dyadic, strain, 181
 Dyadic, stress, 181
 Dye coupling process, 515
 Dyke, W. P., 258
 Dynamic multiplier, 521
 Dynamic plate, resistance, 225
 Dynamic viscosity, 764
 Dynamical state, 446
 DYNAMICS, 178, 398, 679, 753
 continuous, 279
 field, 279
 fluid, 178, 278
 particle, 178
 rigid body, 178
 "Dynamo theories," 394
 Dyne-centimeter, 780
 Dynode, 519

 E region, 348
 Ear, anatomy and physiology of, 309
 Ear drum, 309
 Earth, 45, 50, 293, 344, 535
 atmosphere, 666
 core, 638
 interior temperature, 296
 mantle, 638
 Earth inductor, 396
 Earthquake, 296, 638
 EAS, 138
 Eccentric anomaly, 38
 Echo, 576
 Echo I, 437
 Echo ranging, 665
 Eckert, E. R. G., 320
 Eckert, J. P., 124
 Eclipse, solar, 660
 Ecliptic, 41, 358
 Eddington, Sir Arthur Stanley, 140
 Eddy current, 158, 332, 391, 726
 Edelson, D., 542
 Edge dislocations, 127 (footnote)
 Edge focusing, 153
 Edison, Thomas Alva, 208, 625, 714
 Edison effect, 208, 209, 714
 Edlén, N. E., 150
 Edvac, 124
 Edwards gas density balance, 157
 Effective field method, 399
 Effective jet velocity, 185
 Effective mass, 153
 Effective temperature, 401
 Effective value, 195
 Efficiency, 91
 Efficiency, electric propulsion, 188
 Efficiency, electroluminescence, 203, 204
 Efficiency, machine, 649
 Eickemeyer, 451
 Eigenfunction, 772
 Eigenvalue, 60, 413, 568, 634, 772
 Eikonal matrix, 557
 Einstein, Albert, 81, 86, 101, 139, 251, 314, 325, 531, 771
 Doppler shift, 176
 gravitation, 298
 light scattering, 373
 photoelectric equation, 513
 photon, 263
 radiation emission, 401
 relativity, 264, 404, 408, 419, 428, 619, 712
 Einstein (unit), 439, 509
 Einstein-Bose statistics, 83, 242
 Einsteinium, 735
 Einzel-lens, 212, 213
 Eisner, Leonard, 50
 Ejection velocities, 184
 Ejector pump, 747
 Elastic collisions, 190
 Elastic modulus (polymers), 547
 Elastic scattering, neutron, 604
 Elastic vibration, 761
 Elastic wave, 638
 ELASTICITY, 79, 181, 630
 Electric cell, 200
 Electric charge, 677
 conservation of, 130, 229
 Electric current, 197, 330
 Electric dipole, 159
 Electric field, 322, 327
 Electric field intensity, 550
 Electric flux, 197
 Electric force, 327
 Electric generator, 329
 Electric potential, 550
 Electric potential difference, 197
 ELECTRIC POWER GENERATION, 183
 ELECTRIC PROPULSION, 184, 278
 Electric rocket, 185
 Electric susceptibility, 205
 Electrical conductance, 127
 Electrical conductivity, 127
 Electrical discharge, ultraviolet from, 742
 ELECTRICAL DISCHARGE IN GASES, 189, 202
 Electrical heating, 127, 197
 ELECTRICAL MEASUREMENTS, 193
 Electrical potential, 131
 ELECTRICITY, 196, 534
 Electricity, static, 677
 Electrification, 677
 Electrification, static, 677
 ELECTROACOUSTICS, 16, 199
 Electrochemical cells, 72
 Electrochemical equivalent, 201
 Electrochemical equivalent weight, 346
 Electrochemical potential, 678
 ELECTROCHEMISTRY, 200, 346
 Electrode, 72, 201, 346
 Electrode, hydrogen, 202
 Electrode potential, 201
 Electrodynamical transducer, 199
 Electrodynamical voltmeter, 196
 Electrodynamics, 199, 565
 quantum, 565
 "Electrofax," 515
 ELECTROLUMINESCENCE, 203, 380
 Electrolysis, 200, 346
 laws of, 201
 Electrolyte, 74, 346
 Electrolytic conductor, 200
 Electrolytic rectifier, 610
 Electrolytic solution, 200
 Electromagnetic field, 262

- Electromagnetic force, 317
- Electromagnetic induction, 198, 205, **328**
- Electromagnetic interaction, **228**, 229
- Electromagnetic lens, 211
- Electromagnetic radiation, **583**, 781
- Electromagnetic signal, 576
- Electromagnetic spectrum, **669**, 741
- ELECTROMAGNETIC THEORY, **205**, 385, 426, 493
- Electromagnetic wave, 198, 776
 - propagation of, **554**
- Electromagnetism, 328
- Electromechanical coupling, **199**
- Electrometers, 417
- Electromotive force, 102, 451, 649
 - electrochemistry, 201
 - induced, 198, **327**, 727
- Electromotive series, 74, **200**
- ELECTRON, 57, 84, 117, 172, **208**, 216, 228, 250, 386, 465, 468, 565
 - acceleration of, 704
 - backscattered, 635
 - capture, 597
 - charge to mass ratio, **134**, **208**
 - Compton wavelength, **134**
 - cosmic ray, **138**
 - Dirac theory, 120
 - energy of, 741
 - excited, 237
 - in gases, **189**
 - ionization coefficient, **190**
 - magnetic moment **220**
 - photo-, **512**
 - polarized, **122**
 - positive. *See* Positron
 - in radioactivity, 596
 - radius, **134**
 - rest mass, **133**
 - in semiconductor, 641
 - thermal, 606
 - trapped, 117, 383, 578, 611
 - tunneling, 736
 - wavelength, 211
- Electron affinity, 737
- Electron density, 460
- ELECTRON DIFFRACTION, **210**, 690
 - surface, 690
- Electron emission, 715
- Electron gun, **214**, **217**
- Electron impact, **347**
- Electron lens, **217**
- ELECTRON MICROSCOPE, **211**, 482
 - vacuum system, 215
- Electron number, conservation of, 229
- ELECTRON OPTICS, **216**, 496
- Electron pair production, 604
- Electron paramagnetic resonance, 220, 629
- Electron spectrometer, 227
- ELECTRON SPIN, 51, 59, **219**, 236, 254, 386, 389, 402
 - magnetism, 501
- Electron spin resonance, 220, **387**, 533, 625, 672
- Electron synchrotron, 11
- ELECTRON TUBE, RECEIVING TYPE, **221**
- Electron volt (eV), 9 (footnote), **781**
- Electron-hole pair, **172**
 - bound, **242**
- Electron-phonon interaction, 686
- Electron-positron pair, 583
- Electronic conductivity, **200**, 303, 639
- Electronic instruments, **195**
- Electronic rectifier, 606
- Electronic structure (surfaces), 690
- Electronic transition, **669**
- ELECTRONICS, **226**, 535
- Electro-optic effect, **360**
- Electro-osmosis, 202, **500**
- Electrophoresis, 202
- Electrophotography, 515
- Electrophotoluminescence, **204**
- Electrostatic accelerators, 6
- Electrostatic deflection, **218**
- Electrostatic force, 392
- Electrostatic generator, 9, 667
- Electrostatic lens, 211
- Electrostatic shielding, **417**
- Electrostatic theory, **206**
- Electrostatic thruster, **187**
- Electrostatics, 197
 - in electroacoustics, 199
- Electrostrictive force, 392
- Electrothermal thrusters, **188**
- Electrovalent bond, 444
- ELEMENT, CHEMICAL, 56, 62, 101, **233**, 470
 - abundance in earth's crust, **234**
 - abundance in universe, **234**
 - artificial, **234**
 - cosmic distribution of, **235**
 - definition of, **233**
 - number of, **234**
 - origins of, **235**
 - periodic table, **502**
 - radioactive, **234**, 470
 - terrestrial distribution of, **234**
 - transition, **504**
- Element, transuranium, 462, **732**
- Elemental analysis, 458
- Elementary charge, **133**
- FIFTEENTH PARTICLES, **228**, 455, 470, 476, 565, 778
- Ellipsometry, 612
- Elliptical polarization, **543**
- Elmergreen, G., 408
- Elneutrino, 455
- Elster, Julius, 512, 714
- emf, **327**
 - induced, 726
 - motional, **327**, **328**
 - transformer, **328**
 - variational, **328**
- Emission, avalanche, **204**
- Emission, field, **257**
- Emission, secondary, **635**
- Emission electron microscope, **215**
- Emission spectra, 371, 669
- Emission spectroscopy, 380
- Emissive power, 3
- Emissivity, **337**, 563, **587**
 - spectral, 563
- Emittance, 337, 585, **587**
- Emitter (transistor), 728
- Empedocles, 233
- Emrich, Raymond J., **281**
- Endocrine system, 408
- Endoergic, 469
- Endolymph, 310
- Energy, 313, 455, 647, 780
 - acoustic, 664
 - atomic, **55**
 - bond, **82**
 - capacitor, **89**
 - conservation of, **130**, 419, **781**
 - electrical, measurement of, 193
 - electromagnetic wave, 571
 - electron, 189
 - exchange, 478
 - free, 72, **241**, 291
 - Gibbs, 291
 - Helmholtz, 291

- Energy, (Cont.)**
 internal, 291
 internal, gas, 289
 kinetic, 130, 780
 rotational, 632
 magnetic, 386
 neutron, 456, 457
 nuclear, 55
 particle, 684
 photoelectron, 513
 photon, 522
 plasma, 540
 potential, 130, 550, 780
 radiant, 485
 recoil, 447
 satellite, 36
 solar, 650, 653, 659
 strain, 183
 thermodynamics, 716
Energy band, 154, 611
Energy density (elasticity), 183
Energy equation, 20
Energy gap, 510, 662
 superconductivity, 686
ENERGY LEVEL, 52, 236, 387, 634, 772
 high pressure, 553
 nuclear, 239, 468
 solids, 662
 transitions in, 69
Energy sources, solar, 650
Energy state, excited, 597
Energy transfer, photosynthesis, 525
Energy transfer, radiation, 579
Energy transfer, shock wave, 646
Engine, 270
 air-breathing, 275
 reciprocating piston, 278
 rocket, 270
Engineering, 534
Eniac, 124
Ensemble, 681
 thermodynamics, 719
Enskog, David, 362
Enthalpy, 87, 291, 718
 defect motion, 662
 free, 241
Entrance pupil, 480
ENTROPY, 87, 92, 143, 149, 239, 241, 363, 377, 718
 gas, 291
 metrical, 718
 spectroscopic methods, 672
Eötvös, József, 299
Epitaxy, 721
EPR, 629
Equation of state, 20, 119, 675, 717
Equations of mathematical physics, 262
Equator, 41
EQUILIBRIUM, 182, 240, 262, 377, 506, 510, 7680, 760
 neutral, 241
 phase, 241
 stable, 241
 thermodynamic, 289, 716
 unstable, 24
Equilibrium entropy, 240
Equilibrium state, 242
Equilibrium theory, 363
 chemical, 532
Equivalence, principle of, 621
Equivalent, Coulomb, 201
Equivalent, electrochemical, 201
Equivalent weight, chemical, 201
Eratosthenes, 295
Erg, 780
Error, 415
 systematic, 416
Erway, Duane D., 658
Escape, criterion for, 37
Eta particle, 232
Ettinghausen effect, 306, 619
Euclides, 324
Euclidian space, 139
Euler, Leonhard, 406
Eustachian tube, 310
Eutectic, 507
Evans, Howard T., Jr., 148
Evaporation, 676, 750
Evaporograph, 590
Ewald, P. P., 164
Exchange anisotropy, 391
Exchange charge (fusion), 286
Exchange energy, 82, 478
Exchange forces, 389, 477
Exchange interaction, 28, 256
Excitation, electroluminescence, 203
Excitation, luminescence, 380
Excitation, molecules, 525
Excited electron, 237
Excited state, 59, 117, 458
EXCITON, 117, 242, 400, 664
Excitron, 608
Exclusion principle, Pauli, 250, 256, 475, 477, 502, 533, 715
Exit pupil, 480
Exner, F. M., 86
Exoergic, 469
Exosphere, 296
Expanding universe, 139
Expansion, coefficient of, 244
EXPANSION, THERMAL, 244
Expectation value, 773
Experiment, scientific, 415
Experimental physics, 535, 712
Exploding star, 137
Explosion, chemical, 638
Explosion, nuclear, 638
Exposure, limits, radiation, 306
Exposure meter, 511
Extensive air shower, 138
External inductance, 330
Extinction coefficient, 611
Extinction ratio, Kerr cell, 361
Extraordinary ray, 616
Extreme ultraviolet, 741
Extrinsic material, 172, 640
Eye, 766
Eye piece, 369, 431, 480
Eyring, Henry, 126, 533, 542

F-center, 116
F'-center, 117
F_A-center, 117
F region, 348
Fabry, Charles, 365
Fabry-Perot interferometer, 342, 343
Fabry-Perot resonator, 365
Faceting, 690
Facsimile, 708
Factor, compressibility, 118
Fahrenheit, Gabriel Daniel, 710
Fahrenheit scale, 710
Failla, G., 307, 420
Fall, anode, 192
Fall, cathode, 192
FALLOUT, 247
Far ultraviolet, 741

- Faraday, Michael, 192, **200**, **201**, 205, **248**, 322, 323, 328, **346**, 393, 408, 450
 electromagnetic induction, 205
 electron, **208**
 inductance, 23
 induction, 158
 magnetism, **389**
 Faraday (unit), 201, **346**, 439
 Faraday constant, **134**, 209
 Faraday dark space, **192**
 FARADAY EFFECT, **248**, 555
 Faraday rotation, **248**
 Fast neutron, 456, 458
 Fatigue, **423**
 Feather, Norman, **761**
 FEEDBACK, 148, **249**, 417, 642
 Fenestra, **309**
 Fermat, Pierre, D., **559**
 Fermat's principle, **557**, **559**
 Fermi, Enrico, 81, **250**, **251**
 beta decay, 455
 beta rays, 468
 Fermi function, **475**
 fission, **264**, **265**
 neutrino hypothesis, 778
 theory of beta decay, 209
 Fermi (unit), **142**, **469**
 Fermi distribution, 502
 Fermi energy, 154, 158, **250**, **502**, 658
 Fermi function, **475**, **641**
 Fermi interaction, **779**
 Fermi level, **250**, **251**, 513, 641, 737
 FERMION SURFACE, 128, 154, 155, **251**, 553, 741
 Fermi-Dirac distribution, 83, 128
 Fermi-Dirac statistics, 251, 263, 549, 682
 FERMION-Dirac STATISTICS AND FERMIONS, **250**
 Fermion, 83, **250**, 682
 Fermi-Thomas approximations, 399
 Fermium, **735**
 Ferriaris, G., 451
 FERRIMAGNETISM, **254**, 256, 390, 462
 Ferrite, 391, 462
 Ferroelectric, 246
 Ferroelectric crystal, **255**
 Ferroelectric transducer, 199, 739
 FERROELECTRICITY, **255**
 Ferromagnet, 246
 Ferromagnetic material, 462
 Ferromagnetic rotation, **247**
 FERROMAGNETISM, **256**, 389, 501, 664
 Ferrous metallurgy, 421
 Feynman, Richard P., 263, 400, 455, 546
 Feynman diagrams, 400
 FFAG Accelerators, 12
 Fick, F., **78**, **169**
 Fick's laws (diffusion), **169**, 170
 Field, **261**
 conservative, 754
 depth of, 431
 Dirac, 565
 electric, **197**, 322, 327
 gravitational, **298**
 magnetic, 198, **385**, 392
 Maxwellian, 565
 radiation, 262
 scalar, 753
 vector, 753
 Field dynamics, **279**
 FIELD EMISSION, **258**, 405, 715, 736
 Field equation, **261**
 Field intensity, 327
 electric, 550
 magnetic, 551
 Field ion microscope, **258**
 Field ionization, **258**
 Field method, effective, **399**
 Field of view, **368**
 Field operator, 30
 Field quenching, luminescence, **204**
 FIELD THEORY, 111, **261**
 Film
 photographic, **514**
 superconductive, 723
 thin, 721
 electron diffraction by, 210
 magnetic, 391
 Filter, infrared, 338
 Filter, optical, **432**
 Fine structure, **61**, 70, 220, **238**
 spectroscopic, 219
 Fine structure constant, **134**
 Fingerprinting, molecular, **71**
 Fire-hose instability, **288**
 First law of thermodynamics, 92
 Fissile materials, 55, 267
 Fissile nuclides, 55, **267**
 FISSION, 55, 140, **264**, 350, 355, 457, 470, 471
 fallout, 247
 ternary, **268**
 transport theory, 731
 Fission barrier, **266**
 Fission cross section, **267**
 Fission fragment, **268**, 465
 Fission product, **268**, 475
 Fission product decay, gamma ray, **603**
 Fissionability parameter, **266**
 FitzGerald, George Francis, 426, 428
 Fix (photography), 514
 Fixed array, 576
 Fixed-field alternating-gradient accelerators, 12
 Fixed stars, **178**
 Fizeau, Armand Hippolyte Louis, **757**
 Flare star, 592
 FLIGHT PROPULSION FUNDAMENTALS, **270**
 Floating point representation, **125**
 Flow, Couette, **279**
 Flow, supersonic, **280**
 Fluctuation phenomena, 417
 Fluid, **282**
 Fluid body, **178**
 FLUID DYNAMICS, **178**, **278**
 FLUID STATICS, **281**
 Fluorescence, 365
 resonance, **365**
 sensitized, **381**
 ultraviolet, detection of, 743
 Fluorescence yield, **63**
 Fluorescent lamp, 744
 Fluorescent scattering, 483
 Fluoroscopy, 787
 Flute instability, **287**
 Flutter, 626
 Flutter, aerodynamic, **21**
 Flux, **753**
 density, **388**
 magnetic, **198**, 551
 electric, **197**
 luminous, 379
 magnetic, 183, 328, 330, 726
 radiant, 379
 Flux-gate magnetometer, 397
 Fluxmeter, **396**
 FM, **437**
 Focal length, 481
 Focal point, **368**
 Foci, conjugate, **488**
 Focus, depth of, **482**
 Focus, nuclear, short-range, 56

- Focusing, synchrotron, 705
 Focusing collision, 350
 Fog limit, 126
 Foil, activated, 456
 Fokker-Planck collision term, 541
 Foley, P. J., 380
 Fontana, G., 429
 Foot-lambert, 379, 517
 Foot-pound, 780
 Forbidden gap, 640
 Forbush, 138
 Forbush decreases, 138
 Force, 326, 404, 419, 679, 680
 addition of forces, 680
 buoyant, 754
 centrifugal, 93, 633
 centripetal, 633
 coercive, 257, 390
 composition of, 680
 compressive, 281
 Coriolis, 633
 diffusion drag, 407
 electromotive, 327
 electrostatic, 477
 exchange, 477
 friction, 283
 intermolecular, 291, 378
 Lorentz, 243
 magnetic, 386
 moment of, 327
 ponderomotive, 392
 shear, 281
 tangential, 281
 tensile, 281
 work done by, 780
 Force and acceleration, 180
 Force constants, 83
 Force equation, Lorentz, 205
 Forced convection, 317
 Forging, 422
 Formative lag, 191
 FORTRAN, 125
 Foucault, Jean Bernard Léon, 301, 544, 757
 Fourier, Jean Baptiste Joseph, 315, 454
 Fourier analysis, 321, 454, 775
 Fourier equation, 315
 Fourier series, 22
 Fowler, Alfred, 513
 Fowler, R. H., 240, 258
 Fowler equation, 513
 Fowler-Nordheim theory, 258
 Fowles, Grant R., 634
 Fraction, packing, 476
 Fragment, fission, 268
 Frame of reference, accelerating, 44
 Frank, F. C., 145
 Frank, I. M., 97
 Frank, J., 407
 Frank, O., 406
 Frank dislocation, 144
 Frank-Condon principle, 381
 Franklin, Benjamin, 323
 Fraser, 269
 Fraunhofer, Joseph von, 3, 160
 Fraunhofer diffraction, 161
 Fraunhofer line, 3, 588, 629
 Frautschi, S. C., 112
 Freden, S. C., 578
 Free carriers, electric current, 128
 Free convection, 317
 Free energy, 72, 87, 241, 373, 532
 Gibbs, 291
 Helmholtz, 291, 682
 surface, 689
 Free enthalpy, 241
 Free fall, 44
 acceleration of, 133
 Free gyro, 301
 Free radical, 580
 Freedom, degree of, 506
 Freezing point, 440
 Frenkel, J., 242
 Frenkel excitons, 243
 Frenkel's model, exciton, 242
 "Freon-12," 617
 Frequency, 453, 775, 779
 electromagnetic wave, 571
 generator, 183
 infrared, 337
 Larmor, 387
 light, 370
 natural, 627
 plasma, 248
 power generation, 183
 precessional, 387
 resonance, 628
 Frequency modulation, 437
 Fresnel, Augustin Jean, 160, 426
 interference, 340
 Fresnel coefficient, 722
 Fresnel's law, reflection, 486
 Freundlich, Martin M., 216
 Fricke, H., 420
 FRICTION, 283, 649, 680
 Friction coefficient, 283
 Friedrich, W., 163, 460
 Fringes, interference, 106
 Frisch, Otto, 264
 Fritsch, 451
 Fröhlich, H., 545, 686
 Froome, 759
 Fuchbauer, C., 635
 Fuel, nuclear, 56
 Fuel, propulsion, 278
 Fuel, reactor, 472
 Fuel cell, 75, 658
 Fugacity, 291
 Fulcrum, 648
 Functions, color-matching, 113
 Fundamental mode, 761
 Fundamental physics, 415
 Fundamental series (spectra), 61
 Fundamental tone, 454
 Furth, Harold P., 386
 Fusion, 55, 56, 285, 346, 394, 470, 479, 539
 fallout, 247
 heat of, 312
 nuclear, 285
 shielding, 603
 Fusion curve, 506
 Fusion reactor, 285

 g factor, Lande, 387, 501
 G-parity, 231
 G value, 580
 Gage. *See* Gauge
 Gain, antenna, 25
 Galactic space, 666
 Galaxy, 50, 137, 139
 radio, 593
 Galilean transformation equations, 621
 Galilei, Galileo, 261, 295, 298, 323, 324, 357, 393, 756
 mechanics, 418
 microscope, 429
 sunspots, 344
 Gallium arsenide, 728
 laser, 366

- Galvani, Luigi, 78, **200**
 Galvanic cells, 72
 Galvanoluminescence, **204**
 Galvanometer, **194**
 Gamertsfelder, C. C., 307
 Gamma radiation, 354, 472
 Gamma ray, 120, 264, 266, 447, 465, 467, 581, **596**
 capture, 604
 fission-product-decay, **603**
 in cosmic rays, 137
 inelastic-scattering, 604
 interaction process, 604
 pair production, 549
 prompt, **268**, 458
 prompt fission, **603**
 Gamma transition, 447
 Gamow, George, 596
 Gantmakher effect, 253
 Gap, energy, 510
 Garnet, 254
 Gas, **674**
 density, **157**
 derived properties, 291
 energy functions, 291
 Fermi-Dirac, **250**
 ideal, **289**
 liquefaction, **374**
 Maxwell-Boltzmann, **250**
 perfect, **289**
 permanent, 314
 reference properties, 291
GAS* THERMODYNAMIC PROPERTIES, 291
 Gas constant, **134**, **289**
 Gas dynamics, 20
 Gas equation, **289**
GAS LAWS, 289
 Gas thermometer, **143**, **711**
 Gaseous diffusion (separation), **265**
 Gate (rectifier), 606
 Gauge, McLeod, **748**
 Gauge, Pirani, **748**
 Gauge, redhead, **748**
 Gauge, vacuum, **748**
 Gauge invariance, 131
 Gauge pressure, **281**
 Gauge transformation, **565**
 Gauss, Karl Friedrich, 295, 409, 419
 Gauss' theorem, **197**
 Gaussian error curve, 169
 Gaussian units, 303
 Gaussmeter, 305, **396**
 Gay-Lussac, Joseph, 289, 361
 Gear, **648**
 Gehlen, 408
 Geiger and Marsden experiment, **58**
 Geiger-Müller counter, 370, 465, **566**, 743
 Geitel, Hans Friedrich, 512, 714
 Gell-Mann, M., 112, 400, 455
 General relativity, **620**
 Generator
 electric, **183**, 329
 electrostatic, 9, 677
 magnetohydrodynamic, **184**
 power, isotopic, 57
 Van de Graaff, 9
Geochronology, 295
 Geocorona, 667
GEODESY, 293, 295
 Geodetic parameters, **295**
 Geodimeter, 758
 Geoid, **293**
 Geomagnetic disturbance, 64
 Geomagnetism, **296**, **344**, **348**
 Geometrical acoustics, **16**
 Geometrical optics, **485**
 Geometry, non-Euclidean, 409
 Geophysical prospecting, **297**
 Geophysical Year, International, 295, **344**
 Geophysics, **295**
 George, Barry A., **219**
 Geostrophic wind velocity, **136**
 Geothermal power, 184
 Gerber, H. J., 123
 Gerlach, Walther, **51**
 Germanium, 728
 Germer, Lester Halbert, 163, **210**, 531, 716, 771,
 Getter, 222
 Getter-ion pump, 747
 Ghiorso, A., 734
 Giauque, William Francis, 142
 Gibbs, Josiah Willard, 240, 291, 362, 409, **506**,
 530, 532, **681**, 718, 750
 Gibbs' adsorption equation, **690**
 Gibbs ensemble, 681
 Gibbs function, **718**
 Gibbs phase rule, 719
 Gibson, K. S., 379
 Gimbals, 301
 Ginzburg, V. L., 686
 Giordano, Anthony B., **437**
 Glaciology, **344**
 Glass modulus, 764
 Glass transition temperature, 547, 762, 769
 Glasser, Otto, 420
 Glassy state, **769**
 Global, 589
 Glow, negative, **192**
 Glow discharge, **192**
 Goddard, Robert Hutchings, **188**
 Golay cell, 590
 Gold, T. M. N., 139, 140
 Gold point, 564, 710
 Goldberg, Joshua, 300
 Goldberger, M. L., 112
 Goldhammer, Paul, **476**
 Goldman, D. E., 407
 Goldstein, Eugen, 208, 714
 Goniometer, **147**
 Good, R. H., Jr., **525**
 Goodenough, John B., **158**
 Goodman, Charles D., **141**
 Goodman, Clark, **142**
 Gordon, W., 262
 Gorter, C. J., 142
 Goudsmit, Samuel A., 209, **219**, 236, 790
 Gouy, M., 86
 Governor, 249
 Gradient, 754
 Grain (rocket), **273**
 Grain boundary, 661
 Grain size, thin film, 722
 Gram molecular weight, 438
 Gramme, Zénobe Théophile, 450
 Grashof number, 317
 Grassmann, H., 751
 Grating, diffraction, 494
 Gravimetry, **294**, **344**
 Gravitation, 79, 293, **297**
 Gravitational constant, 133
 Gravitational potential, **298**
 Gravitational theory, 659
 Gravity, 44, 67, 293, 404, 419
 Gravity, center of, 680
 Gravity, specific, **156**
 Gray, J. A., 120
 Gray, L. H., 420
 Gray body, 337, **587**

- Greek absorption bands, 117
 Green, George, 450
 Green, M. S., 683
 Green's function, 207, 570
 Greenhouse effect, 592
 Gregg, Donald C., 102
 Gregg, S. J., 20
 Grid, 222
 Grimaldi, Francesco M., 160, 340
 Groseclose, B. Clark, 550
 Ground state, 59, 238
 Ground waves, 436
 Group, isospin, 231
 Group theory, 261
 Group velocity, 756, 776
 Growth (condensation), 126
 Growth spiral, 127
 Grubbé, 306
 Gruneisen relation, 245
 Gudden, B., 204
 Gudden and Pohl effect, 204
 Guest molecules, 243
 Guggenheim, E. A., 81
 GUIDANCE, INERTIAL, 333
 Guidance requirements, space flight, 38
 Guinier scattering, x-rays, 786
 Guinn, Vincent P., 460
 Gurney, R. W., 596
 Gutenberg, B., 295
 Gwinn, Cecil W., 77
 Gyro, 333, 556
 Gyro, rate, 336
 Gyrocompass, 301
 Gyromagnetic ratio, 219, 389
 Gyromagnetic ratio of proton, 134
 Gyromagnetic resonance, 254
 GYROSCOPE, 301, 335
 Gyroscope, atomic, 485

H-center, 117
 "H-sheet," 544
H-theorem, Boltzmann, 363
 Haddock, F. T., 591
 Hagen-Rubens relation, 613
 Hagstrom, H. D., 637
 Hahn, O., 264
 Haldane, John Burden Sanderson, 406
 Hale, George Henry, 393
 Half cell, 72
 Half-life, neutron, 456
 Half-life (radioactivity), 598
 Hall, Edwin Herbert, 248, 303, 397
 Hall angle, 304
 Hall coefficient, 303, 546
 Hall coefficient, semiconductor, 641
 Hall coefficient factor, 304
 Hall devices, 305
 Hall effect, 129, 248, 393
 HALL EFFECT AND RELATED PHENOMENA, 303
 Hall field, 303
 Hall mobility, 304, 546
 Hall-effect device, 397
 Hallwachs, Wilhelm, 512
 Hamer, Walter J., 203
 Hamilton, Sir William Rowan, 419, 751
 optics, 558
 Hamiltonian, 634, 681, 773
 Hamiltonian operator, 565
 Hamilton's principle, 559
 Handling of information, 148
 Hanford, Washington, 265
 Hanson, A. O., 123
 Harasima, Akira, 694
 Hardening, dispersion, 522
 Harder, D. S., 249
 Harmonic generator, 254
 Harmonic motion, 760
 Harmonic motion (simple), 454
 Harmonic oscillation, 627
 Harmonic series, 761
 Harris, Forest K., 196
 Harrison, Roland H., 293
 Harrison, W. A., 252
 Hart, J. C., 307
 Hart, R. W., 373
 Harteck, Paul, 539
 Hartkemeier, Harry P., 68
 Hartley, 499
 Hartmann number, 392
 Hartree-Fock theory, 399
 Harvey, B. G., 736
 Häuy, René Just, 147
 Haynes, Sherwood K., 64
 HAYSTACK facility, 437
 Heading, John, 412
 HEALTH PHYSICS, 306, 421
 HEARING, 78, 309
 Hearon, J. Z., 408
 HEAT, 87, 311, 534, 780
 Joule, 128
 mechanical equivalent of, 313
 quantity, 311
 radiation, 585
 specific, 244
 HEAT CAPACITY, 87, 292, 312, 313
 Debye theory, 166
 spectroscopic methods, 672
 Heat death, universe, 364
 Heat equation, 262
 Heat exchanger, 375
 Heat of combustion, 87
 Heat of compression, 119
 Heat of fusion, 312
 Heat of sublimation, 750
 Heat of vaporization, 292, 312, 750
 Heat pump, 374, 617
 HEAT TRANSFER, 315
 Heat transfer, radiant, 588
 Heater, 222
 Heating, aerodynamic, 21
 Heating, electrical, 127, 197
 Heat-treatment, 422
 Heaviside, Oliver, 409, 751
 Heidt, Lawrence J., 510
 Heisenberg, Werner, 58, 109, 209, 210, 216, 263, 320, 325, 477, 574, 713
 HEISENBERG UNCERTAINTY PRINCIPLE, 51, 320
 Heitler, W., 82
 Helicon, 664
 Heliosphere, 348
 Helium, 377, 468
 in superfluidity, 688
 Helium laser, 366
 Hellwarth, R. W., 367
 Helmholtz, Hermann Ludwig Ferdinand von, 78, 208, 291, 313, 325, 420, 453, 718
 Helmholtz free energy, 682, 718
 Helmholtz function, 718
 Helmholtz's theorem, 393
 Hemispheric long wave, 424
 Henry, J., 331, 450
 Heraclitus, 233
 Herapath, John, 362
 Hering theory of vision, 116
 Hermitian matrix, 410, 413
 Hermitian operator, 566
 Herriott, D. R., 366
 Herschel, Sir William, 585

- Hertz, Heinrich Rudolph, 206, 208, **512**, 576
 Hertz potential, 206
 Hertzprung-Russell diagram, **50**
 Herzberger, M. J., 616
 Hess, Victor, **137**
 Hess, Wilmot N., **579**
 Hetényi, M., **512**
 High current discharges, **193**
 HIGH-VOLTAGE RESEARCH, **322**
 Hilbert, David, 409
 Hill, A. V., 78, 79
 Hill, T. L., **126**
 Hiroshima, Japan, **265**
 Hiss (with aurora), 65
 HISTORY OF PHYSICS, **323**
 Hittorf, Johann Wilhelm, **200**, 714
 Hodgkin, A. L., 407
 Hogerton, John F., **57**
 Hohraum, 585
 Hohmann transfer orbit, **39**, 48
 Hole, **641**
 trapped, 117
 Holland, M. G., **509**
 Hollerith, Herman, **124**
 Hologram, **514**
 Holt, Charles, **330**
 Homogeneous medium, 529
 Homopolar bond, 444
 Hooke, Robert, 182
 microscope, **429**
 polarization, 543
 Hookean solid, 630
 Hooke's law (elasticity), **182**
 Hopkins, Robert F., **370**
 Hopping electron, 546
 Hopping mechanism, 642
 Horn antenna, **27**
 Horn equation, 529
 Horsepower, 780
 Hosemann, Rolf, **168**
 Host crystal, 243
 Hot carriers, **204**
 Hot working, 422
 Houston, W. V., **59**
 Hoyle and Narlikar theory, **140**
 Hubble, Edwin Powell, 139, 300
 Hubble's constant, 139
 Huckel, E., 201
 Huijter, Karel, **325**
 Hume-Rothery, W., **422**
 Humidity, relative, 676
 Humphreys, Curtis J., **71**
 Hund, F., **256**, 389
 Hund's rule, 28, **256**, **389**
 Hunt, F. L., 785
 Hunt, Frederick V., 93
 Hunting, 148
 Hutchison, T. S., 741
 Huxley, A. F., 407
 Huygens, Christian, **340**, 429, **493**, **615**
 polarization, 543
 Huygens' hypothesis, **340**
 Huygens' principle, 160, 493
 Hydrodynamics, 20, **278**
 Hydrogen, 461
 Hydrogen atom, 387
 Hydrogen bomb, **265**, 470
 Hydrogen corona, 667
 Hydrogen electrode, **202**
 Hydrogen ion concentration, 202
 Hydrology, **297**
 Hydro-magnetic wave, **348**
 Hydromagnetics, 540
 Hydrometer, 157
 Hydrophone, 666
 Hydrothermal growth, **146**
 Hylleraas, E. A., 773
 Hyman, 352
 Hypercharge, **29**, 229, 684
 Hyperfine line, 484
 Hyperfine structure, 52, **62**, **238**
 Hyperon, **229**
 Hysteresis, 254, **257**, 331, 390

 I. T., 597
 Ice age, 297
 Iceland spar, 496
 Ice-point, 710
 Ideal gas, 91, **289**, **314**, 675, 715
 Ideal gas flow, **280**
 Ignitor, **608**
 Ignitron, **608**
 triggered spark gap, 562
 IGY, 295, **344**
 Illuminance, 516
 Illumination
 dark ground, **483**
 in microscopes, **483**
 phase contrast, **483**
 Illuminometer, **516**
 Image, 480
 virtual, 487
 Image force, **224**, 737
 Image formation, **487**
 Image intensifiers, 227
 Imbibition, **500**
 Imes, E. S., 68, 71
 Impedance, 23, **103**
 complex, **24**
 terminal, **25**
 Impedance bridge, **194**
 Impedance matching, transducers, **199**
 Impedance tube, 529
 Implosion, **265**
 Impulse, 47, **326**, 753
 angular, **327**
 linear, **326**
 specific, 185
 IMPULSE AND MOMENTUM, **326**
 laws of, **326**
 Impurity diffusion, **170**
 Impurity level, semiconductor, 641
 Impurity, semiconductors, 128
 Incandescent lamp, 714
 Inch, **133**
 Incidence, angle of, **486**
 Inclined plane, 647, 648
 Incus, **309**
 Independence, charge, **231**
 Independent particle model, 474, 560
 Index, extraordinary ray, 360
 Index, ordinary ray, 360
 Index of refraction, 360, 368, 373, **486**, 495, 611
 Indicating instruments, **195**
 Indifferent point, 507
 Indirect semiconductor, 641
 INDUCED ELECTROMOTIVE FORCE, 198, **327**, 726, 727
 INDUCTANCE, 23, **330**, 726, 727
 external, **330**
 Faraday's law, 23
 internal, **330**
 measurement of, **193**
 mutual, **330**
 parallel, **331**
 self, **330**
 series, **331**
 Inductance coil, **331**
 Induction, of charge, **677**

- Induction, electromagnetic, 198
- Induction, magnetic, 385, 388
- Inductor, 331
 - measurement of, 193
- Inelastic scattering, gamma ray, 604
- Inelastic scattering, neutron, 604
- Inertia, 403
 - moment of, 71
- Inertial frame of reference, 135, 621
- INERTIAL GUIDANCE, 333
- Inertial navigation, 301
- Inertial observer, 633
- Information handling, 148
- Information theory, 409
- Infrared, 370, 586
 - detector, 338
 - military applications, 339
 - polarization, 543
- INFRARED RADIATION, 337
- Infrared spectra, 83, 339, 670
- Infrared spectroscopy, 83, 339, 670
- Infrasonic, 17
- Ingalls, Robert L., 450
- Ingot, 422
- Injection electroluminescence, 203
- Injection velocity, 37
- Injector, rocket, 272
- Inner product, vectors, 752
- Inner transition element, 504
- Inorganic chemistry, 102
- Input (feedback), 249
- Instability
 - electrical discharge, 193
 - fire-hose, 288
 - flute, 287
 - interchange, 287
 - mirror, 288
 - velocity space, 287
- Instrument transformers, 196
- Instruments
 - direct-current, 195
 - electronic, 195
 - indicating, 195
 - nuclear, 464
 - optical, 480
 - rectifier, 195
 - thermocouple, 195
- Insulation, electrical, 323
- Insulator, 128, 129, 239, 254
- Integrated circuit, 144
- Integrating sphere, 517
- Integrator, electric, 124
- Integrator, mechanical, 124
- Intensity, reflection coefficient, 612
- Intensity, sound, 32, 453, 463, 528
- Intensity, wave, 775
- Interaction, electromagnetic, 229
- Interaction, strong, 29, 229, 683
- Interaction, weak, 778
- Interaction process, gamma, 604
- Interaction process, neutron, 604
- Interatomic distances, 71
- Interband transition, 611
- Interchange instability, 287
- Interfacial angles, 147
- Interference, 106, 321, 340, 493, 512
 - constructive, 493
 - destructive, 493
- INTERFERENCE AND INTERFEROMETRY, 340
- Interference microscopy, 483
- Interferometer, 342, 426
 - Fabry-Perot, 342, 343
 - Michelson, 342
- Interferometry, 340
- Intergalactic matter, 139
- Intergalactic space, 666
- Intermediate neutron, 456
- Intermolecular forces, 381
- Internal energy, gas, 289
- Internal energy function, 717
- Internal inductance, 330
- Internal state quantum numbers, 110
- Internal stresses, 181
- Internal symmetry, 231
- International Commission on Illumination, 113, 516
- International Geophysical Cooperation, 295
- International Geophysical Year and International Years of the Quiet Sun, 295, 344
- International Polar Year, 344
- International Years of the Quiet Sun, 344
- Internuclear distance, 672
- Interplanetary flight, 39, 188
- Interplanetary space, 666
- Interplanetary travel, 39, 188
- Interstellar material, 50
- Interstitial, 171, 350, 661
- Intraband transition, 611
- Intramolecular forces, 381
- Intrinsic (field theory), 261
- Intrinsic material, 172
- Intrinsic semiconductor, 640
- Invariance, 31, 228
 - gauge, 131
 - method of, 732
 - time reversal, 131
- Invariant imbedding, 732
- Inverse-square law, 357
 - illuminance, 516
- Inversion, combined, 229
- Inversion, population, 485
- Inversion, space, 229
- Ion, 200, 346
 - positive, 191, 405
 - radiation chemistry, 580
- Ion bombardment, 635
- Ion current, positive, 191
- Ion pair, 580
- Ion rocket, 48
- Ion source, synchrotron, 704
- Ion-acoustical wave, 348
- Ionic bond, 81
- Ionic cloud, 201
- Ionic polarizability, 159
- IONIZATION, 53, 190, 238, 346, 468, 539
 - by radiation, 580
 - secondary, 190
- Ionization chamber, 456, 465, 743
- Ionization coefficient, electron, 190
- Ionization energy, 504
- Ionization potential, 346
- Ionization theory, 440
- Ionizing radiation, 78, 581
- Ionophone, 199
- IONOSPHERE, 65, 296, 344, 348
- Ionospheric currents, 65
- IPTS, 564
- IQSY, 295, 344
- Iris, 766
- Irradiance, solar, 653
- IRRADIATION, DISPLACED ATOM, 350
- Irreversibility, 363
- Isobar, 696
- Isochromatic, 512
- Isoclinic, 512
- Isolating circuit, 305
- Isolator, 254
- Isomer, nuclear, 597
- Isomeric transition, 458, 597

- Isomerism, **101**
 Isospin group, **231**
 Isospin multiplets, **231**
 Isotactic polymer, **547**
 Isotherm, adsorption, **18**
 Isothermal process, **91**
 ISOTOPE, **62, 71, 234, 352, 458, 475, 696**
 tracer, **594**
 Isotope effect, superconductivity, **686**
 Isotope processing, **355**
 Isotope shift, **62, 559**
 Isotopic power generators, **57**
 Isotopic spin, **231**
 Isotropic crystal, **616**
 Isotropy, in cosmic rays, **137**
 Ives, H. F., **178**
 Ivey, Henry F., **204**
- J**-operator, **24**
 Jacobi, C. G. J., **419**
 Jacobi identity, **752**
 Jacquez, A., **408**
 Jakobi, William W., **75**
 Jammer, Max, **680**
 Jansky, K. G., **591**
 Janssen, Zacharias, **429**
 Javan, A., **366**
 Jaynes, E. T., **240**
 Jeffreys, Sir Harold, **295**
 Jeffries, Z., **476**
 Jelley, J. V., **97**
 Jet stream, **426**
 Jet velocity, **185**
 Jodrell Bank, **591**
 Joenk, R. J., **257**
 Joffe bars, **287**
 Johannson, H., **212**
 Johnsen, Russell H., **236**
 Johnson, Francis S., **668**
 Johnson, J. B., **716**
 Joining (metals), **422**
 Joliot-Curie, Frédéric, (originally, Joliot, Frédéric), **468, 596**
 Joliot-Curie, Irene, **596**
 Jones, H., **422**
 Jones, R. S., **123**
 Jordan, **140**
 Jordanus de Nemore (Nemorarius), **679**
 Josephs, Jess J., **454**
 Joule, James Prescott, **291, 313, 331**
 Joule heat, **128, 720**
 Joule (unit), **780**
 Joule-Thomson expansion, **142, 291**
 Joule-Thomson liquefier, **375**
 Judd, Deane B., **116**
 Junction, planar, **601**
 Junction, *p-n*, **172**
 Junction transistor, **728**
 Jupiter, **535**
- K**-matrix, **400**
k-space, **640**
 Kabir, P., **779**
 Kaibaba, R., **408**
 Kammerlingh Onnes, Heike, **142, 375**
 Kant, Immanuel, **325**
 Kaon, **229**
 Kastler, A., **483**
 Kaula, William M., **294**
 Kayser (unit), **337**
 Kekulé, Friedrich, **101**
 Kelvin, Lord (William Thomson), **325, 710, 720**
 skin effect, **650**
 Kelvin bridge, **194**
 Kelvin relations, thermoelectricity, **720**
 Kelvin scale, **710**
 Kennedy, J. W., **733**
 Kennelly, A. E., **650**
 Kenney, Robert W., **86**
 Kepler, Johannes, **297, 324, 357, 429**
 Kepler's equation, orbital position, **38**
 KEPLER'S LAWS OF PLANETARY MOTION, **45, 297, 357, 386**
 Kerr, John, **360, 367**
 Kerr cell, **361, 367, 757**
 KERR EFFECT, **174, 360**
 Kerr shutter (laser), **367**
 Kerst, Donald W., **76, 704**
 Kieffer, William F., **439**
 Kilocalorie, **312**
 Kilogram, **133**
 Kiloton (explosives), **264**
 Kinematics, **178, 752**
 Compton effect, **120**
 Kinetic energy, **130, 780**
 elasticity, **183**
 fission, **269**
 in molecule, **238**
 rotational, **632**
 Kinetic friction coefficient, **283**
 KINETIC THEORY, **51, 57, 361, 540, 712**
 Kinetic theory of gases, **51, 57**
 Kinetics, **178**
 chemical, **98**
 Newton's laws of motion, **180**
 King, Allen L., **406, 619**
 King, Gerald W., **447**
 Kingslake, Rudolph, **617**
 Kingston, R. H., **367**
 Kirchhoff, Gustave R., **160**
 thermal radiation, **318**
 Kirchhoff's formula, **529**
 Kirchhoff's integral (diffraction), **1**
 Kirchhoff's law
 electricity, **23**
 heat, **319**
 thermal radiation, **3, 585**
 Kirkendall effect, **171**
 Kirkwood, **542**
 Klein, F., **262**
 Klein, O., **120**
 Klein and Nishina equation, **121**
 Klein and Nishina theory, **120**
 Klein-Gordon equation, **30, 262**
 Kliauga, Paul J., **703**
 Klystron, **226**
 Knipping, P., **163, 460**
 Knock-on, **350**
 Knoll, M., **212**
 Kohbrausch, Hans, **78**
 Kohler, **432**
 Kohlrausch, Rudolf Hermann Arndt, **757**
 Kolmogoroff, **424**
 Kosevich, A. M., **156**
 Kossel, Walter, **63, 82, 127, 444**
 Kostitzin, V. A., **406**
 Kostkowski, Henry J., **564**
 Krakatoa, **296**
 Kremers, Howard E., **603**
 Krishnan, K. S., **599**
 Kronecker symbol, **413**
 Kronig-Penney model, **692**
 Kubo, R., **683**
 Kuhn, H. G., **62**
 Kuiper, G. P., **539**
 Kumar, Kailash, **400**
 Kunzler, J. E., **89**
 Kuper, C. G., **547**

- Kurşunoglu, Behram, **575**
 Kurti, **142**
 Kusch, P., **52, 219**

 Lag, formative, **191**
 Lag, statistical, **191**
 Lagemann, Robert T., **776**
 Lagrange, Joseph Louis, **419**
 Lagrange equations, **419**
 Lagrangian, **565**
 Laing, Ronald A., **502**
 Lamb, Sir Horace, **739**
 Lamb, W. E., **52, 447**
 Lamb shift, **51, 61, 263, 569**
 Lamb wave, **739**
 Lambda anomaly, **315**
 Lambda hyperon, **229**
 Lambda particle, **232**
 Lambda point, **84**
 Lambert, Heinrich Johann, 3(footnote), **379, 517, 527, 613**
 Lambert cosine law of reflection, **519, 527, 588, 613**
 Lamé, Gabriel, **182**
 Lamé's constants, **182**
 Lamp, fluorescent, **744**
 Lamp, incandescent, **714**
 Land, E. H., **515**
 Landahl, H. D., **407**
 Landau, L. D., **155, 158, 502, 686, 779**
 Landau damping, **541**
 Landau levels, **156, 158**
 Landau self-trapping, **546**
 Landau's principle, **779**
 Landé, Alfred, **790**
 Landé *g*-factor, **387, 501, 629**
 Landsøerg, H. E., **297**
 Lane, C. T., **689**
 Langevin, Paul, **390, 502**
 magnetic moment, **175**
 Langevin function, **501**
 Langevin theory, **390**
 Langmuir, Irving, **539, 715**
 Langmuir oscillation, **541**
 Lanthanide contraction, **601**
 Laplace, Pierre Simon de, **206, 325, 409**
 Laplace equation, **206, 262, 551**
 Laplace operator, **262**
 Laplace transform, **643**
 Larmor, Sir Joseph, **248, 387, 426, 790**
 Larmor frequency, **248, 387, 388, 556, 629, 790**
 Larmor's theorem, **158**
 Lasarew, **502**
 LASER, **365, 371, 485, 543**
 photography, **514**
 Latent heat, **750**
 Latent image, **514**
 Latimer, R. M., **736**
 Latitude, astronomic, **294**
 Lattice, crystal, **147, 378, 399, 783**
 Lattice defects (semiconductors), **128**
 Lattice imperfection, **260**
 Lattice structure, **244**
 Lattice vibration, **128, 663**
 Laue, Max von, **163, 460, 783, 784**
 diffraction theory, **165**
 Laue method, **783**
 Laughlin, J. S., **421**
 Lauritsen, C. C., **258**
 Laval, **165**
 Laval spots, **165**
 Lavoisier, Antoine Laurent, **101, 233**
 Law
 of areas, **360**
 of atmospheres, **86**
 of conservation, **130, 229**
 angular momentum, **130**
 energy, **130**
 momentum, **130**
 of electrolysis, **201**
 of thermodynamics, **717**
 Lawrence, Ernest O., **150, 704**
 Lawrencium, **736**
 Laws, F. A., **650**
 Layard, Sir Austin Henry, **429**
 LCAO, **446**
 Leaning Tower of Pisa, **298**
 Le Châtelier's principle, **241**
 Lecher system, **757**
 Leclanché cell, **72**
 Leduc, **306**
 Lee, Reuben, **727**
 Lee, T.-D. **131, 455, 546, 779**
 Leeuwenhoek, Anton van, **429**
 Lee-Yang theory, **144**
 Le Fèvre, R. J. W., **175**
 Leibnitz, Gottfried Wilhelm, **124**
 Lemaitre, Georges Édouard, Abbé, **140**
 Lemm, **168**
 Lenard, Philipp, **208, 512, 635**
 Lennard-Jones, J. F., **290**
 Lennard-Jones potential, **290**
 LENS, **367, 429, 480**
 achromatic, **431**
 apochromatic, **431**
 compensating, **431**
 concave, **492**
 condenser, **431**
 convex, **492**
 crystalline, **766**
 defects in, **492**
 Einzel, **213**
 electromagnetic, **211**
 electron, **217**
 electrostatic, **211**
 fluorite, **431**
 Huygenian, **431**
 negative, **367**
 periplane, **431**
 positive, **367**
 reflection by, **491**
 unipotential, **213**
 widefield, **431**
 Lens testing, **369**
 Lensless photography, **514**
 Lenz, Heinrich, **330**
 Lenz's law, **158, 330, 389**
 Lepton, **229**
 Level, energy, **236**
 Level-crossing effect, **485**
 Lever, **647**
 Leverrier, Urbain, **325**
 Levin, A. A., **71**
 Levine, Joseph, **106**
 Lewis, Gilbert Newton, **82, 444, 533**
 Libby, W. F., **138**
 Lifetime, particle, **230**
 Lifshitz, I. M., **156**
 LIGHT, **228, 370, 534**
 absorption of, **49**
 emission of, **49**
 measurement of, **516**
 modulation, **485**
 polarized, **483, 542**
 speed of, **133, 755**
 velocity of, **133, 755**
 wave theory of, **494**
 Light propagation, law of, **620**

- Light quanta, 771
- Light ray, 485
- LIGHT SCATTERING, 372, 443
 - polymers, 547
- Light sources, 371
- Lightning, 322
- Lilienfeld, J. E., 258
- Lilley, A. E., 591
- Limit of resolution, 481
- Limiter, 254
- Linac, 6, 9
- Linde liquefier, 375
- Linde process, 375
- Linden, Bernard R., 522
- Lindsay, Robert, 629
- Line broadening, 373
- Line defect, 661
- Line of force, 198, 385
- Line spectra, 59
- Linear accelerator, 6, 9
- Linear combination of atomic orbitals, 446
- Linear energy transfer, 583
- Linear impulse, 326
- Linear momentum, 326, 455
- Liouville, Joseph, 681
- Liouville equation, 681
- Lionville theorem, 241
- Lippincott, Ellis R., 672
- Lippmann, B., 109, 111
- LIQUEFACTION OF GASES, 374
- Liquid, 376, 674
 - density, 157
- Liquid drop model, nucleus, 478
- Liquid phase, 506
- Liquid propellant, 272
- LIQUID STATE, 376
- Liquid structure, 378
- Liquidus, 507
- Liter, 133
- Littauer, Raphael M., 563
- Livingston, M. Stanley, 150, 704
- Llewellyn, F. B., 105
- Lloyd, Humphrey, 340
- Local fallout, 247
- Local vertical system, 335
- Locomotion, human, 406
- Lodestone, 535
- Loeb, A. L., 322
- Loferski, Joseph J., 528
- Lofgren, Edward J., 153
- London, F., 82, 144, 686
 - superconductivity, 686
- Long range force, 477
- Long-chain molecules, 547
- Longitude, astronomic, 294
- Longitudinal wave, 739, 775
- Lorentz, Hendrik Antoon, 57, 262, 303, 389, 392, 422, 426, 428, 621, 790
 - electromagnetic waves, 554
 - harmonic oscillator, 613
 - Zeeman effect, 209
- Lorentz equations, 621
- Lorentz force, 129, 150, 243, 303, 389, 392
- Lorentz force equation, 155, 205
- Lorentz invariance, 621
 - scattering, 110
- Lorentz transformations, 262
- Lorentz trapping, 287
- Lorentz triplet, 62, 790
- Lorentz unit, 790
- Lorentz-Lorenz formula, 124, 723
- Los Alamos, New Mexico, 265
- Loschmidt, Joseph, 362
- Loss angle, 159
- Loss compliance, 762
- Loss factor, capacitance, 90
- Loss modulus, 762
- Loss tangent, 159, 762
- Lossev, 203
- Lotka, Alfred J., 406
- Loudness, 17, 309, 310, 463
- Loudness level, 463
- Lovibond color systems, 115
- Low, E., 546
- Low pressure area, 136
- Lubricant, 284
- Lufburrow, Robert A., 92
- Lukens, H. R., Jr., 595
- Lumen, 379, 516
- LUMINANCE, 115, 379, 516
- LUMINESCENCE, 117, 203, 365, 380
- Luminosity, sun, 659
- Luminous efficiency, 379
- Luminous flux, 379, 516
- Luminous intensity, 379, 516
- LUNAR FLIGHT, 39
- Lung, 408
- Lykoudis, Paul S., 395
- Lyman, Theodore, 60, 742
- Lyman series, 60, 742
- Lyman-alpha radiation, 668, 742
- Lyon, David N., 376
- M-center, 117
- Macelwane, J. B., 295
- Mach, Ernst, 140, 409, 419
 - gravity, 404
- Mach number, 275, 646
- Mach principle, 140
- MACHINE, SIMPLE, 647
- Machurek, Joseph E., 479
- Mackenzie, John D., 770
- Mac Rae, Alfred U., 211
- Macroscopic theory, 712
- Magic number, 269, 474
- Magnet, 256, 391
 - superconducting, 144, 687
- Magnetic bottle, 56
- Magnetic current, 328
- Magnetic deflection, 218
- Magnetic dipole, 387
- Magnetic energy, 386
- Magnetic field, 75, 198, 256, 303, 385, 387, 388, 392, 397
 - cyclotron resonance, 153
 - earth, 668
 - internal, 450
- Magnetic field intensity, 551
- Magnetic film, 724
- Magnetic flux, 183, 198, 327, 328, 330, 551
- Magnetic flux density, 198, 327, 551
- Magnetic force, 386
- Magnetic induction, 385, 388
- Magnetic lens, 212
- Magnetic materials, 385
- Magnetic moment, 256, 257, 386, 388, 395, 456, 501
 - atomic, 389
 - nuclear, 62
- Magnetic potential, 551
- Magnetic properties, 175
- Magnetic pumping, plasma, 286
- Magnetic quantum number, 503
- Magnetic resonance, 51, 386, 624, 629, 664
 - accelerators, 150
 - nuclear, 169
- Magnetic storms, 65, 296, 344

- Magnetic susceptibility, 158, 206
- Magnetic tape, 625
- Magnetic vector, potential, 206
- MAGNETISM, 256, 385, 388, 534, 636
- Magnetization, 175, 256, 388, 395
 - spontaneous, 256
- Magnetization curve, 256
- Magnetoacoustic effect, 253
- Magnetoconductivity, 129, 306
- Magnetocrystalline anisotropy, 390
- Magnetofluid dynamics, 278
- MAGNETO-FLUID-MECHANICS, 391, 540, 556
- Magneto-gas-dynamics, 392
- Magneto-hydrodynamic generator, 184
- Magneto-hydrodynamics, 392, 540, 556
- Magneto-hydro-mechanics, 392
- Magneto-mechanical ratio, 249
- Magnetometer, 395, 485
- MAGNETOMETRY, 395
- Magnetomotive force, 451
- Magneton, 629
 - Bohr, 134, 175, 219, 386, 389
 - nuclear, 134, 387
- Magneto-optic effect, 361
- Magneto-oscillation, 155
- Magnetopause, 348
- Magneto-plasma-dynamics, 392
- Magnetoresistance, 129, 253
 - Corbino, 306
 - longitudinal, 306
 - transverse, 305
- Magnetoresistivity tensor, 129
- Magnetosphere, 667
- Magnetostatic theory, 206
- MAGNETOSTRICTION, 246, 397
- Magnetostrictive force, 392
- Magnetostrictive materials, 199
- Magnetostrictive transducer, 739
- Magnetothermoelectric refrigerator, 619
- Magnetron, 226, 436
- Magnification, 429, 480
 - lens, 490
- Magnifier, 369
- Magnifying power, 480
- Magnon, 664
- Maiman, T. H., 365
- Main sequence star, 50, 652
- Major groups, elements, 506
- Majority carrier, 172, 609
- Majority charge carriers, 203
- Maleus, 309
- Malus, Étienne Louis, 495, 615
- Malus effect, 495
- Malus' law, 544
- Mandelstam representation, 112
- Manhattan Project, 265, 421
- Manometer, 282, 442
- Mantle (earth), 638
- MANY-BODY PROBLEM, 363, 398, 540
- Many-particle problem, 363, 398, 540
- Maran, Stephen P., 594
- Mariner II, 667
- Mars, 358, 535
- Marshak, R. E., 455
- Martin, John W., 423
- Martinville, Léon Scott de, 625
- Marton, L., 514
- Marx method, 322
- MASER, 365, 400, 485
- Maser oscillator, 403
- Mass, 44, 326, 404
 - center of, 680
 - critical, 140
 - particle, 684
 - relativistic, 326
 - stopping power, 583
- Mass action, law of, 241
- Mass analyzer, 404
- Mass and energy, equivalence of, 620
- MASS AND INERTIA, 403
- Mass defect, 477
- Mass number, 234, 458
- Mass spectrometer, 352, 404
- Mass spectrometry, 101
- Mass spectroscopy, 404, 671
- Mass unit, 133, 469
- Mass-to-charge ratio, 404
- Master equation, plasma, 683
- Master equation (pumping), 484
- Material, fissile, 55
- Material, interstellar, 50
- Materials, magnetic, 385
- Mathematical biology, 406
- Mathematical biophysics, 406
- Mathematical physics, 262, 408, 535
 - equations of, 262
- MATRICES, 409
- Matrix, 409
- Matrix algebra, 544
- Matrix elements, 412
- MATRIX MECHANICS, 412, 771
- Matrix theory, Heisenberg, 531
- Matter, 565, 674
 - constituents of, 231
- Matthews, P. T., 685
- Matthiessen's rule, 128, 723
- Mauchly, John, 124
- Mauro, A., 407
- Maximum thermometer, 711
- Maxwell, James Clerk, 81, 102, 205, 250, 328, 385, 392, 408, 426, 532, 554
 - aether, 619
 - optics, 493
 - viscosity, 362
- Maxwell particles, 251
- Maxwell relation (dielectrics), 159
- Maxwell stress, 392
- Maxwell theory, 208, 263
- Maxwellian distribution, 80, 242, 362, 456
- Maxwell's equations, 143, 262, 370, 385, 392, 522, 540, 572
 - light scattering, 372
- Maxwell-Boltzmann distribution, 83
- Maxwell-Boltzmann statistics, 242, 250
- Maxwell-Faraday law, 328
- Maxwell-Wagner effect, 160
- Mayer, Joseph E., 290, 379, 407, 682
- Mayneord, W. V., 420
- McCleod gauge, 748
- McClung, F. J., 367
- McClure, J. W., 129
- McCready, L. L., 591
- McFee, Raymond H., 339
- McGrath, J. W., 100
- McIntyre, D. P., 426
- McKenzie, A. E. E., 483
- McMillan, E. M., 151, 704, 733
- Mean anomaly, 38
- Mean free path, 731
- Mean free path, charge carriers, 128
- Measurements, electrical, 193
- MEASUREMENTS, PRINCIPLES OF, 415
- Meatus, 309
- Mechanical advantage, 647
- Mechanical energy, conservation of, 183
- Mechanical equivalent of heat, 313
- Mechanical integrators, 124
- Mechanical rectifier, 610

- Mechanical vibrations, 416
 MECHANICS, 67, **418**, 534, 621, 679
 elasticity, 181
 matrix, **412**
 resonance, 627
 statistical, **680**
 wave, **771**
 MEDICAL PHYSICS, **420**
 Medium, homogeneous, 529
 Meissner effect, **158**, **685**
 Meitner, Lise, **264**
 Mel, **453**
 Mel, Howard C., **585**
 Melting point, congruent, 507
 Meltzer, Carl H., **226**
 Melvin, Mael A., **31**
 Membrane, **500**
 Membrane potential, 407
 Mendeleev, Dmitri, **502**
 Mercalli scale, 638
 Mercury, 535
 Mercury arc, ultraviolet from, **742**
 Mercury barometer, **282**
 Mercury discharge tube, 589
 Mercury vapor tubes, **224**
 Mercury-arc rectifier, **608**
 Meson, 76, 229, 231, 455, 478, 560, 779
 Mesopause, **426**
 Mesosphere, **296**, **426**
 Mesotron. *See* Meson
 Metal, 252, 256, **421**
 resistivity, 128
 Metallic bond, **81**
 METALLURGY, **421**
 Metamagnetism, **391**
 Metastable ion, 405
 METEOROLOGY, 296, 344, **423**
 Meter, 132, 343
 Metrical entropy, 718
 Metzger, F., 122
 Meyer, C. F., 71
 Meyer, Lothar, **502**
 Meyer, Victor, 441
 Meyerhofer, Dietrich, **738**
 Mhos per meter, **127**
 Michelson, Albert A., 107, 325, **426**, **757**
 interference, 342, 494
 Michelson interferometer, **342**, **426**
 MICHELSON-MORLEY EXPERIMENT, 343, **426**, 620, 712
 Microminiaturization, 77
 Microphone, 199
 Microradiography, x-ray, 787
 MICROSCOPE, **429**, 480
 electron, **211**
 electron emission, **215**
 field emission, **258**
 field ion, **258**
 mirror, **215**
 objective lens, **369**
 polarizing, 545
 scanning, **215**
 shadow, **215**
 stereoscopic, **432**
 thermionic emission, **215**
 Microscopic theory, 712
 Microscopy, interference, **483**
 Microtron, **10**
 Microwave, 576
 Microwave amplification, 400
 Microwave devices, 249, 254
 MICROWAVE SPECTROSCOPY, **433**, 533, 672
 Microwave telegraphy, 707
 MICROWAVE TRANSMISSION, **436**
 Microwave tube, 226
 Middle ultraviolet, **741**
 Midnight vector, **358**
 Miles, John L., **144**
 Milky way, **50**
 Millikan, Robert Andrews, 57, 132, **208**, **258**, **742**
 Millikan's oil drop experiment, **209**
 Mills cross, 591
 Milton, 269
 Minimum thermometer, 711
 Minkowski space, 408
 Minority carrier, 172, 610
 Minority carrier injection, 642
 Minority charge carriers, 203
 Mirror, 485
 defects in, 492
 metal, 722
 microscope, 481
 parabolic, 490
 Mirror instability, **288**
 Mirror inversion, **131**
 Mirror microscope, **215**
 Mirror symmetry, **131**
 "Misch" metal, 602
 Missiles, 67
 MO (molecular orbital), 446
 Mobility, carrier, **641**
 Mobility, charge carrier, **128**, **304**
 Mode, normal, **507**
 Moderator, 141, 457, 459
 neutron, **473**
 Modern physics, 535
 MODULATION, **437**
 in telemetry, 709
 light, 485
 Raman effect, 599
 Modulus, elastic (polymers), 547
 Moeckel, W. E., **189**
 Mohler, Orren C., **661**
 Mohole, **296**
 Mohorovicic discontinuity, **296**
 Molar quantities, 439
 Molar solution, 439
 Molar volume, 439
 Mole, **438**
 MOLE CONCENTRATION, **438**
 Molecular beam, **51**, **54**
 Molecular biophysics, **79**
 Molecular complex, 445
 Molecular compound, 445
 Molecular orbital, 446
 method, **82**
 Molecular polarizability, **159**, 541, 600
 Molecular spectra, 68, **672**
 Molecular spectroscopy, **68**, 672
 Molecular structure, 444, 533, 542
 Molecular theory, 749
 gases, **118**
 Molecular transition, 401
 Molecular vibration frequency, **600**
 Molecular weight, 94, **439**, **445**
 of polymers, 547
 Molecular weight determination, 765
 Molecule, 51, 57, 80, 100, 440, **444**, 541, 676
 diatomic, **68**, **69**
 long-chain, 547
 non-polar, 173
 polar, **173**, **541**
 MOLECULES AND MOLECULAR STRUCTURE, 444
 Moment, 680
 quadrupole, 62
 Moment of force, **327**, 752
 Moment of inertia, 71, 759
 Moment of momentum, 566

- Momentum, 76, 110, **326**, 571, 681, 753
 - angular, 237, **327**, **480**, **632**
 - atomic, 60
 - conservation, **130**, **326**
 - linear, **326**
 - photon, 522
- Momentum space, 640
- Monitoring (radiation), 308
- Monochromatic vision, **113**
- Monolayer, atomic, 721
- Monomer, **548**
- Monomolecular layers (condensation), 127
- Monopole (acoustics), 529
- Monopole, antenna, **26**
- Monopropellant, **273**
- Monte Carlo method, **732**
- Montgomery, D. J., **679**
- Morgan, Karl Z., 307, **309**
- Morgan, R., 734
- Morley, Edward Williams, 343, **426**
- Morphology, crystal, **147**
- Moseley, Henry Gwyn-Jeffreys, 233, 785
- Mössbauer, Rudolf L., 229, **447**
- MÖSSBAUER EFFECT, **447**
- Mossotti (polarization), **174**
- Motion, Brownian, **86**
- Motion, planetary, **357**
- Motion, wave, **775**
- Motional emf, **327**, **328**
- MOTORS, ELECTRIC, **450**
- Mott, N. F., 422
- Motz, J. W., 122
- Mount Palomar microscope, **482**
- Moving-iron instruments, **196**
- Mueller, Hans, **544**
- Mullen, Albert A., **250**
- Müller, E. W., **258**, **260**
- Müller, J. H., 124
- Multifunction tubes, **226**
- Multiplets, isospin, **231**
- Multiplexing, 709
- Multiplication constant, **266**
- Multiplicative, conservation law, 229
- Multiplier, 305
- Multivibrator, 561
- Mu-meson, in cosmic rays, **137**
- Mu-neutrino, 455
- Munsell color systems, **115**
- Muon, 229
- Muon capture, 779
- Muon decay, 779
- Muon number, conservation of, 229
- Music, 35, **453**
- MUSICAL SOUND, **453**
- Mutscheller, A., 307
- Mutual inductance, **330**
- Muzzle velocity, 67
- Myers, Raymond R., **631**

- Nachtrieb, Norman H., **170**
- Nadeau, Gérard, **183**
- Nagasaki, Japan, **265**
- Narlikar, 140
- Narrow-band transformer, 727
- Natural convection, **317**
- Natural frequency, 627
- Natural radioactivity, **596**
- Navier-Stokes equations, 317
- Navigation, 335
 - radar, 576
- Near ultraviolet, **741**
- Nebula, **50**
- Néel temperature, **28**, **390**
- Negative glow, **192**
- Negative lens, **367**
- Negative temperature, **401**
- Negatron, 597
- Neodymium, 366
- Neon laser, 366
- Neptune, 535
- Neptunium, **733**
- Nernst effect, **306**
- Nernst glower, **589**
- Nerve, **77**
- Nerve excitation, 407
- Nerve fiber, 407
- Nervous system, 408
- Neuberger, Jacob, **246**
- Neumann, Franz Ernst, **330**
- Neumann equation, **330**, **332**
- Neuron, 77, 79
- Neutrino, 228, **455**, 468, 469, 778
 - fission, 268
 - in cosmic rays, **138**
 - in radioactivity, 596
 - solar, **651**
- Neutron, 55, 264, 266, 350, 354, 399, **456**, 458, 460, 465, 470, 471, 474, 476, 584
 - bombardment by, 733
 - delayed, 266, 475, 478, **603**
 - fast, 456, 458
 - fission, 141
 - interaction process, 604
 - intermediate, 456
 - prompt, **268**, 478, **603**
 - resonance, 456
 - slow, **456**
 - thermal, **456**, 458, 584
- NEUTRON ACTIVATION ANALYSIS, **458**
- Neutron attenuation, **605**
- NEUTRON DIFFRACTION, **460**
- Neutron diffusion, 141
- Neutron flux, 458
- Neutron inelastic scattering, 458
- Neutron rest mass, **134**
- Newlands, **502**
- Newton, Sir Isaac, 2, 261, 295, 323, 357, **403**, 409, **418**
 - diffraction, **340**
 - gravitation, **297**
 - mechanics, **418**
- Newtonian liquid, 630
- Newtonian mechanics, **180**
- Newton-meter, 780
- Newton's equation (heat), **317**
- Newton's experiment (prism), 495
- Newton's laws of motion, 136, **180**, 261, 270, 326
- Newton's rings, **342**
- Nicholson, William, **200**
- Nicol, William, 544
- Nicol prism, 544
- Nielsen, Lawrence E., **548**
- Nightglow, **65**
- Nijboer, B. R. A., **1**
- Nishina, Y., 120
- Nit, **379**
- Nitrobenzene, 361
- NMR, 629
- Nobelium, 736
- Nodal point, **368**
- Node, 761
- Noise, 31, **416**, 453, **463**
- NOISE, ACOUSTICAL, **463**
- Noise, maser, 400
- Noise, radar, **577**
- Noise, sonar, 666
- Noise abatement, 464
- Nomenclature, **694**

- Nonequilibrium mechanics, 683
- Non-Euclidean geometry, 409
- Non-ferrous metallurgy, 422
- Noninteracting particles, 399
- Nonlinear circuit, 104
- Non-Newtonian system, 765
- Non-polar molecules, 173
- Nordheim, L. W., 258
- Normal atmosphere, 133
- Normal dispersion curve, 3
- Normal mode, 507
- Normal mode analysis, 399
- Normalization of wave function, 773
- Noys, 18
- Nozzle, rocket, 272
- n*-type material, 172
- Nuclear debris in cosmic rays, 138
- Nuclear disintegrations, 76, 457
- Nuclear energy, 55, 470, 651
- Nuclear explosion, 638
- Nuclear fuel, 56
- Nuclear fusion, 285
- NUCLEAR INSTRUMENTS, 464
- Nuclear magnetic moment, 62, 672
- Nuclear magnetic resonance, 169, 387, 533, 629
- Nuclear magneton, 134, 387
- Nuclear moments, 62, 143, 672
- Nuclear power, 265, 479
- Nuclear precession magnetometer, 397
- Nuclear radiation, 344, 464, 467
- NUCLEAR REACTION, 50, 468
- NUCLEAR REACTOR, 55, 196, 247, 307, 354, 457, 459, 471, 478
- Nuclear spin, 52, 62
- Nuclear spin resonance, 533
- Nuclear stability, 476
- Nuclear steam plants, 184
- NUCLEAR STRUCTURE, 55, 471, 474
- Nuclear track, 467
- Nuclear transmutation, 476
- Nuclear weapon debris, 247
- Nucleation, 126, 144, 722
- Nucleogenesis, 50
- Nucleon, 474, 560
- Nucleon resonance, 232
- NUCLEONICS, 476
- Nucleus, 58, 228, 264, 386, 398, 456, 474, 476, 565
 - compound, 470
 - in electricity, 196
 - protons in, 559
 - residual, 470
 - tunneling in, 738
- Nuclide, 55, 696
 - fissile, 267
 - stability, 596
- Null matrix, 409
- Number, magic, 269, 474
- Number, mass, 234
- Number, quantum, 220, 236, 502
- Numerical analysis, 125
- Numerical aperture, 430, 481
- Nusselt number, 317
- Nyquist, 149
- Oak Ridge, Tennessee, 265
- Oberman, Carl, 541
- Oberth, H., 188
- Objective
 - camera, 369
 - lens, 431
 - microscope, 369
 - telescope, 369
- Oboe, 758
- Obry, 301
- Observability, 574
- Ocean, acoustic waves in, 665
- Ocean currents, Coriolis effect, 136
- Oceanography, 297, 344
- Oersted, 388
- Ohm, Georg Simon, 385
- Ohm (unit), 127
 - determination of, 193
- Ohm's law, 198, 385
 - alternating currents, 22
- Oil immersion objective, 481
- Olbers, Heinrich Wilhelm Matthäus, 139
- Oliver, Jack, 638
- Olson, Harry F., 464
- Olson, John M., 526
- Omega resonance, 232
- Omegatron, 749
- One-dimensional box, 772
- O'Neil, Stephen J., 644
- Onnes, Kammerlingh, Heike, 375
- Onsager, L., 155, 159, 542
- Onsager equation, 159
- Open capillary technique (diffusion), 169
- Opera house, acoustics of, 35
- Operational calculus, 409
- Operator, 263
- Oppenheimer, J. Robert, (fission), 265
- Optic axis, 496, 616
- Optic nerve, 767
- Optical activity, 496
- Optical axis, lens, 367
- Optical branch, phonon, 507
- Optical coating, 722
- Optical crystallography, 147
- Optical dielectric constant, 159
- OPTICAL INSTRUMENTS, 480
- Optical path, 493
- Optical phonon, 641
- OPTICAL PUMPING, 402, 483
- Optical pyrometry, 563
- Optical rotation, 248
- Optical transition, 381
- Optics, electron, 216
- OPTICS, GEOMETRICAL, 370, 485, 557
- Optics, nonlinear, 372
- OPTICS, PHYSICAL, 370, 493
- Optics, physiological, 370
- Orbit, 358
- Orbital, 503
- Orbital angular momentum, 237
- Orbital magnetic quantum number, 503
- Orbital moment, 175
- Orbital quantum number, 503
- Order (spectrum), 494
- Order parameter, 687
- Ordinary ray, 496, 616
- Organ of Corti, 310
- Organic chemistry, 102
- Organizations, physics, 535
- Orthochromatic film, 515
- Orthohydrogen, 560
- Ortho-positronium, 549
- Oscillation, 759
 - betatron, 75
 - dipole, 207
 - harmonic, 627
 - plasma, 539
- Oscillator, blocking, 561
- Oscillator, harmonic, molecular, 69
- Oscillator, mechanical, 627
- OSCILLOSCOPE, 496
- Osgood, T. H., 742
- Osmometry, 442
- OSMOSIS, 498

- Osmotic pressure, 442, 499, 530
 polymers, 547
 Ossicle, 309
 Ostwald, Wilhelm, 530
 color systems, 115
 Ostwald's dilution, 530
 Otto cycle, 277
 Outer product, vectors, 752
 Output (feedback), 249
 Oval window, 310
 Overman, Ralph T., 598
 Overpotential, 202
 Overshoot, oscilloscope, 497
 Overtone, 454, 761
 Overvoltage, 202
 Oxidation, 346, 580
 electrochemistry, 201
 metals, 423
 photosynthesis, 509
 Oxide-coated cathode, 223
 Ozone, 297

 p-type material, 172
 Pacinotti, Antonio, 450
 Packing fraction, 476
 Page, Thornton, 51
 Pair production, 121, 465, 549, 583, 584
 Pairing force, 475
 Panchromatic film, 515
 Panoramogram, parallax, 514
 Parabolic mirror, 490
 Parabolic-reflector antenna, 27
 Paracrystal, 166
 Paraelectric, 255
 Parahydrogen, 560
 Parallax shift, 42
 Parallax panoramogram, 514
 Parallel inductance, 331
 Parallelogram law, 326, 751
 Parallelogram rule, 326, 751
 Paramagnetic material, 389
 Paramagnetic rotation, 249
 PARAMAGNETISM, 256, 389, 402, 501
 Parametric amplification, 254, 418
 Para-positronium, 549
 Paraxial ray, 367
 Pardue, L. A., 307
 Parity, 62, 131, 144, 229, 524
 Parker, E. N., 667
 Parker, H. M., 307
 Parkins, William E., 239, 781
 Parrant, G., Jr., 107
 Parsec, 139 (footnote)
 Partial pressure, 750
 measurement of, 749
 Particle
 charged, 465, 582
 elementary, 228
 energetic, 137
 lifetime of, 230
 Maxwell, 251
 spin orientation, 111
 strange, 229
 Particle accelerator, 150
 Particle conjugation, 229
 Particle dynamics, 178
 Particle precipitation into atmosphere, 65
 Particle transport, 730
 Partition function, 81, 126, 682, 719
 Parvin, Richard H., 40, 336
 Pascal, Blaisé, 124, 281
 Pascal's law, 281
 Pascal, P., 211
 susceptibility, 175

 Paschen, F., 60, 191, 220, 790
 Paschen series, 60
 Paschen-Back effect, 220, 791
 Paschen's law, 191
 Passive network, 104
 Path, optical, 493
 Patlak, C., 407
 Pauli, Wolfgang, 61, 209, 236, 251, 455, 475, 477, 502, 559
 Pauli exclusion principle, 61, 83, 220, 236, 250, 251, 256, 475, 477, 502
 Pauli matrices, 568
 Pauling, Linus, 533, 542
 Pawsey, J. L., 591
 Payne-Scott, R., 591
 Pearson, F., 757
 Pease, Francis Gladheim, 757
 Peierls, R., 155
 Pekar, S. I., 546
 Peltier, J. C. A., 720
 Peltier effect, 618, 720
 Pelzer, H., 545
 Pendulum, 759, 780
 Pendulum magnetometer, 396
 Penetration depth, superconductivity, 685
 Pentagrid converter, 226
 Pentode, 223, 225
 Perfect gas, 119, 289
 Periapysis distance, 37
 Pericenter, 46, 47
 Perifocus, 46
 Perigee, 359
 Perihelion, 357
 Perilymph, 310
 Period, 775
 vibration, 759
 PERIODIC LAW AND PERIODIC TABLE, 502
 Periodic potential, lattice, 252
 Periodic table, 61, 502, 505
 Periodic time, 579
 Periscope, 369
 Peritectic, 507
 "Permalloy" film, 724
 Permanent gas, 374
 Permeability, 198, 254, 303, 332
 Permittivity, 128, 159, 205, 550, 551
 relative, 159
 Perot, A., 365
 Perovskites, 361
 Perrin, Jean Baptiste, 86, 208
 Persistent current, 688
 Persistent internal polarization, 515
 Perturbation, orbital, 39
 Perturbation theory, 570
 Perveance, 217, 224
 Peter, Martin, 54
 Peters, Robert W., 311
 Phase (of matter), 506, 675, 775
 under high pressure, 553
 Phase angle, 23, 331
 Phase angle, oscillator, 628
 Phase change, reflection, 621
 Phase contrast, microscope, 432
 Phase contrast illumination, 483
 Phase diagram, 377
 Phase diagram (metallurgy), 422
 Phase equilibrium, 241
 Phase modulation, 438
 Phase plane, 81
 PHASE RULE, 506, 719
 Phase shifter, 254
 Phase space, 681
 Phase stability, principle of, 8
 Phase stability (synchrocyclotron), 152

- Phase velocity, 776
 Phasotrons, 152
 Phi resonance, 232
 Phillips, Norman E., 315
 Phoenix mirror experiment, 287
 Phon, 453, 463
 Phonautograph, 625
 Phonograph, 625
 PHONON, 400, 463, 507, 641, 664, 688, 741
 absorption, 611
 dispersion of, 663
 in superconductivity, 686
 Phonon spectrum, 507, 508
 Phonon-electron interaction in amplifiers, 200
 Phosphor, 203, 383
 Phosphorescence, 204, 381
 Photoabsorption, 346
 Photocathode, 519
 Photocathode, ultraviolet, 743
 Photocell, 338, 418
 PHOTOCHEMISTRY, 509, 525, 579
 PHOTOCONDUCTIVITY, 338, 370, 510, 512
 semiconductors, 129
 PHOTOELASTICITY, 496, 512, 545
 Photoelectric cell, 370
 Photoelectric effect, 121, 371, 583
 gamma, 604
 Photoelectric emission, 258
 Photoelectric pyrometer, 564
 PHOTOELECTRICITY, 512
 Photoelectron, *os*, 637
 Photoemission, 519
 PHOTOGRAPHY, 514
 lensless, 514
 Photoionization, 346, 526
 Photoluminescence, 380
 Photometer, 516
 Photometric quantities, 518
 PHOTOMETRY, 516
 PHOTOMULTIPLIER, 97, 227, 466, 513, 519, 564, 637
 Photomultiplier, ultraviolet, 743
 PHOTON, 228, 447, 468, 469, 509, 522, 543, 565, 571, 771
 electrical discharge, 192
 in cosmic rays, 137
 polarization of, 122
 thermal radiation, 586
 virtual, 228
 Photon energy, 522
 Photon gas, 84
 Photon momentum, 522
 Photon propulsion, 278
 Photon rocket engine, 278
 Photopeak, 460
 Photopigments, 115
 Photoplasticity, 512
 Photoreceptor, 766
 Photosphere, 344, 394, 660
 PHOTOSYNTHESIS, 525, 658
 Photoviscoelasticity, 512
 PHOTOVOLTALIC EFFECT, 511, 524, 526, 656
 PHYSICAL ACOUSTICS, 12, 528
 PHYSICAL CHEMISTRY, 102, 530
 Physical metallurgy, 422
 Physical optics, 493
 Physical quantities, 694
 PHYSICS, 534
 atomic, 57
 experimental, 535
 fundamental, 415
 history of, 523
 mathematical, 408, 535
 modern, 535
 radiological, 421
 solar, 659
 solid-state, 535, 661
 surface, 689
 theoretical, 535, 712
 trends in, 535
 Physics organizations, 535
 Physiological acoustics, 17
 Physiology, 420
 Picard, E., 570
 Picht, Johannes, 1
 Picht-Luneberg integrals, 1
 Pickup reaction, 471
 Pictet, Raoul, 142
 Pierce, P. H., 650
 Piezoelectric transducer, 739
 Piezoelectricity, 199
 Piezomagnetism, 199
 Piezoresistance, 199
 Pigment, 115, 525
 Pike, Julian M., 136
 Pile (nuclear reactor), 265, 471, 478
 Pile, voltaic, 200
 Pi-meson, in cosmic rays, 137
 Pinajian, J. J., 355
 Pinch effect, 193, 394
 Pines, David, 546
 Pion, 228, 229
 Pipe Reynolds number, 280
 Pippard, A. B., 252, 686
 Pirani, M., 748
 Pirani gauge, 748
 Pisa, leaning tower, 298
 Pitch, 309, 453, 463
 Pitot tube, 281
 Pizii, Hippolyte, 450
 Placzek, G., 599
 Planar junction, 610
 Planar transistor, 729
 Planck, Max, 49, 101, 209, 236, 319, 325, 571, 771
 thermal radiation, 586
 Planck's constant, 60, 69, 83, 98, 134, 216, 251, 319, 571
 Planck's equation, 319
 Planck's theory, 209
 Plane of polarization, 248
 Planet, 56, 357, 535, 666
 motion of, 41, 357, 418
 radiation from, 537
 PLANETARY ATMOSPHERES, 71, 535
 Planetary orbit, 45
 PLASMA, 192, 346, 348, 392, 539, 682
 fusion, 285
 Plasma frequency, 248, 556
 Plasma oscillations, 399
 Plasma propulsion, 395
 Plasma thruster, 187
 Plasma wave, 556
 Plasmon, 664
 Plate, deflection, 497
 Plate current, 224
 "Plowshare," 265
 Plücker, Julius, 208
 Pluto, 539
 Plutonium, 55, 57, 265, 472
p-n junction, 172, 527, 642
 Pneumatic detector, 590
 Pohl, R., 204
 Poincaré, Jules Henri, 299, 364, 409, 419, 426, 428, 545
 Poincaré sphere, 545
 Poincaré's theorem, 399
 Point, focal, 368
 Point, nodal, 368
 Point, principal, 368

- Point defect, 661
- Point group, 147
- Point-contact transistor, 728
- Poise, 764
- Poiseuille, Jean Louis Marie, 78, 420, 764
- Poiseuille flow, 280
- Poison, reactor, 473
- Poisson, Siméon D., 182, 206, 298, 419, 550
- Poisson bracket, 681
- Poisson's equation, 206, 298, 550
- Poisson's ratio, 182, 762
- "Polacolor," 514
- POLAR MOLECULES, 173, 541
- Polar Year, International, 344
- Polarimeter, 496
- Polarizability, 541
 - molecular, 600
- Polarization, 255
 - circular, 543
 - dielectric, 205
 - elliptical, 543
 - galvanic cell, 73
 - light, 370, 495
- Polarization field, 242
- Polarization of photons, 122
- Polarization vector, 97, 174, 555
- Polarization vector, particles, 111
- Polarized electrons, 122
- POLARIZED LIGHT, 483, 512, 542, 616
- Polarizer, 512, 543
- Polarizing microscopes, 432, 545
- "Polaroid," 514
- POLARON, 154, 400, 545, 664
- Pole, magnetic, 388
- Polonium, 456
- Polymer, 547
 - molecular weight, 440
- POLYMER PHYSICS, 547
- Pomerantz, Martin A., 345
- Pompage optique, 483
- Ponderomotive force, 392
- Pope, Alan, 21
- Population inversion, 366, 401, 485
- Portis, Alan M., 249
- Position vector, 179
- Positive column, 192
- Positive electron. *See* Positron
- Positive ion, 191
- Positive ion (solutions), 201
- Positive ion current, 190
- Positive lens, 367
- Positive ray, 352, 404
- POSITRON, 210, 228, 468, 549, 565
- Positron annihilation, 253
- Positronium, 549
- POTENTIAL, 386, 550
 - breakdown, 191
 - chemical, 241
 - electrical, 131
 - electrochemical, 678
 - electrode, 201
 - Hertz, 206
 - intermolecular, 290
 - magnetic, 551
 - membrane, 407
 - periodic, 663
 - lattice, 252
 - standard electrode, 202
- Potential barrier, 737
- Potential difference, 191
 - electric, 197
- Potential energy, 130, 780
 - in molecule, 238
- Potential equation, 262
- Potential gradient, 303
- Potential gradient, electric, 550
- Potential scattering, 109
- Potential well, 737
- Potentiometer, 193
- Pound, 133
- Pound, G. M., 127
- Pound, R. V., 169, 299, 448
- Powder diffraction analysis, 788
- Powder metallurgy, 422
- Powder model, nucleus, 474
- Powell, C. F., 126
- Power, 647, 780
 - dispersive, 495
 - electrical, 24, 331
 - measurement of, 193
 - geothermal, 184
 - lens, 493
 - magnifying, 480
 - nuclear, 479
 - resolving, 481
 - sound, 32
 - thermoelectric, 720
 - transmission of, 726
- Power factor, 24, 331
 - capacitors, 90
 - polymers, 548
- Power generation systems (electric propulsion), 185
- Power supply, electron microscope, 215
- Poynting, John Henry, 207
- Poynting vector, 207, 557
- Prandtl number, 317
- Prebuncher, 8
- Precession, 753
- Precessional frequency, 387
- Precipitation, 146
- Precipitation hardening, 422
- Present, R. D., 291
- Pressure, 67, 279, 281, 552, 675, 745
- Pressure, absolute, 281
- Pressure, atmospheric, 281
- Pressure, barometric, 282
 - gas, 291
 - gauge, 281
 - osmotic, 442, 499
 - vapor, 749
- PRESSURE, VERY HIGH, 552
- Pressure area, 136
- Pressure broadening, 59
- Pressure gradient, 754
- Pressure microphone, 32
- Preston's rule, 790
- Prévost, Pierre, 318
- Prévost's principle, 318
- Price, William J., 467
- Primary cell, 74
- Primary radiation, cosmic rays, 137
- Principal point, 368
- Principal quantum number, 531
- Principal series (spectra), 61
- Principia*, 324, 419
- Principle of equivalence (mass), 299
- Principle of superposition, 280
- Prins, I. A., 164
- Prism, 487, 615
 - Ahrens, 544
 - dispersion, 494
 - Foucault, 544
 - Newton's experiment, 495
 - Nicol, 544
 - Wollaston, 544
- Probability density, 773
- Process metallurgy, 421
- Procyon, 358

- Programming, 124
 Projectile velocity, 67
 Projection lantern, 480
 "Prompt critical" condition, 266
 Prompt fission gamma ray, 603
 Prompt gamma ray, 268, 458, 597
 Prompt neutron, 268, 478, 603
 PROPAGATION OF ELECTROMAGNETIC WAVES, 554
 Propellant, liquid, 272
 Propellant, solid, 273
 Proper motions, stars, 42
 Properties, extensive, 291
 Properties, intensive, 291
 Proportional counter, 465, 466
 Propulsion, 270
 electric, 184, 278
 photon, 278
 rocket, 47
 Prospecting, geophysical, 297
 PROTON, 399, 456, 465, 474, 475, 476, 559
 acceleration of, 704
 Compton wavelength, 134
 conversion, 779
 gyromagnetic ratio of, 134
 in cosmic rays, 137
 in electricity, 196
 in radioactivity, 596
 nuclear decay, 549
 solar, 668
 strong interaction, 683
 trapped, 578
 Proton moment, 134
 Proton rest mass, 134
 Proton synchrotron, 11
 Protonosphere, 348
 Proton-proton chain, 651
 Psychological acoustics, 17
 Ptolemy, 324
 Puchstein, A. F., 451
 Pulley, 647
 Pulley, differential, 647
 PULSE GENERATION, 561
 Pulse inversion, maser, 401
 Pulse transformer, 727
 Pulse waveform, 561
 Pump
 diffusion, 747
 ejector, 747
 getter-ion, 747
 rotary, 745
 sorption, 747
 sputter-ion, 747
 vacuum, 745
 vapor, 745
 Pumping, 401
 cryogenic, 748
 optical, 483
 Pupil, entrance, 480
 Pupil, exit, 480
 Purcell, E. M., 169
 Pycnometer, 157
 Pyrometer, 564, 710, 711
 photoelectric, 564
 total radiation, 564
 two color, 564
 PYROMETRY, OPTICAL, 563

 Q-factor, 331, 402
 Q-switched mode, 367
 Quadrupole moment, 62
 Quality (sound), 454
 Quantities, physical, 694
 Quantization, second, 263
 Quantization, superconductivity, 144

 Quantum, 371, 507, 534
 magnetic flux, 134
 Quantum concept, 101
 Quantum detector, 590
 QUANTUM ELECTRODYNAMICS, 61, 231, 399, 565
 Quantum field theory, 30, 399
 relativistic, 30, 565
 Quantum liquids, 377
 Quantum mechanics, 262, 320, 712, 771
 Quantum number, 60, 220, 229, 236, 502, 531
 Quantum states, symbols, 696
 Quantum statistical mechanics, 682
 Quantum statistics, 83
 QUANTUM THEORY, 570
 Quasar, 300, 593
 Quasiparticle, 440, 664
 nucleus, 399
 Quasi-stellar object, 300, 593
 Quasi-stellar sources, 300, 593
 Quimby, Edith, 420

 R-center, 117
 Rabi, I. I., 52
 Rabinowicz, Ernest, 285
 Rad, 421
 RADAR, 421, 576, 758
 measurement, 577
 Radar equation, 576
 Radial acceleration, 632
 Radial distribution function, molecules, 119
 Radiance, 113, 379, 586
 spectral, 563
 Radiant energy, 485
 Radiant flux, 379
 Radiant intensity, 379
 Radiation, 565, 581, 771
 Cerenkov, 96
 deceleration, 204
 delayed, 472
 gamma, 472
 heat transfer, 315, 318
 infrared, 337
 intensity, 25 (footnote)
 IONIZING, BASIC INTERACTIONS, 581
 nuclear, 467
 resonance, 483
 synchrotron, 137
 THERMAL, 88, 570, 585
 ultraviolet, 741
 RADIATION BELTS, 138, 539, 577, 667
 Van Allen, 138
 RADIATION CHEMISTRY, 579
 Radiation constant, 134, 586
 Radiation damage, 350
 Radiation field, 262
 Radiation length, 84
 Radiation pattern, antenna, 25
 Radiation pressure, 528
 Radiation resistance, 581
 Radiation slide rule, 587
 Radiative loss, 582
 Radiative transfer, 730
 RADIO ASTRONOMY, 591
 Radio aurora, 65
 Radio galaxy, 593
 Radio star, 592
 Radio telemetry, 708
 Radio telescope, 50, 591
 Radio wave, polarization, 543
 Radio wave propagation, 348
 Radioactive decay, 55, 476
 RADIOACTIVE TRACERS, 594
 Radioactive waste, 307

- RADIOACTIVITY**, 306, 455, 458, 467, 468, **595**, 732,
778
 artificial, **354**, **468**
 fallout, 247
Radiocarbon dating, 138
Radioisotope, **352**, **354**, **472**, **594**
Radioisotope cell, 186
Radiological physics, 421
Radiolysis, 580
Radiometer, 590
Radiometric quantities, 518
Radiometric spectrometer, 590
Radiometry, 590
Radionuclides (fallout), 247
Radiosonde, 424
Radium, 75, 307
Radius, atomic, **504**
Radius vector, 45
Raimes, Stanley, 774
Rain, 106
Rajaraman, R., 404
Rall, W., 408
Raman, Sir Chandrasekhara, 599
Raman effect, 70, 380
RAMAN EFFECT AND RAMAN SPECTROSCOPY, **599**
Raman scattering, 366
Raman spectroscopy, 671
Raman spectrum, 82, 373
Ramjet, 278
Ramsey, N. F., 456
Random error, 415
Random flights (diffusion), 171
Range, beta ray, 468
Range, radar, 576
Ranger spacecraft, 657
Rankine, William John Macquorn, 710
Rankine cycle, 186
Rankine scale, 710
Raoult, François Marie, 440, **750**
Raoult's law, **750**
RARE EARTHS, 250, **601**
Rashevsky, 79, **406**, **408**
Ratio, magneto-mechanical, 249
Rauch, Lawrence L., 709
Ray, gamma, 447, **596**
Ray, light, 485
Rays, cosmic, 137
Rayleigh, Lord James (John William Strutt), 373,
586, 739
 refractometer, 341
Rayleigh (unit), 64
Rayleigh number, 317, 394
Rayleigh scattering, 167, 380, 599
Rayleigh wave, 739
Rayleigh-Gans scattering, 373
Rayleigh-Jeans equation, 586
Reactance, 23, 331
Reaction
 chain, 55, 56
 deuterium-deuterium, 56
 deuterium-tritium, 56
 nuclear, 458, 469
 pickup, 471
 stripping, 471
 thermonuclear, 56
 transfer, 471
Reaction coordinate, 533
Reaction kinetics, 98
Reaction rate, chemical, 533
Reaction rate theory, 170
Reactor, 604
 breeder, 734
 fusion, **285**
 nuclear, 186, 247, **265**, 307, **471**
 Reactor fuel, **472**
 REACTOR SHIELDING, **603**
Réaumur, René Antoine Ferchault de, 710
Réaumur scale, 710
Rebka, G. A., 229, 448
Receptor cells, 77
Reciprocating piston engine, 278
Recoil electron, 120
Recoil energy, 447
Recoilless gamma transitions, 447
Recombination, 641
Recombination, charge carriers, 510
Recombination, ionospheric, **349**
Recombination, ions, 466, 584
Recombination center, 511
Recombination electroluminescence, 149
Recording, sound, 625
Recrystallization, 146
Rectangular cell, 74
Rectification (electric), **606**
RECTIFIER, 305, **606**
 controlled, 562, 606
Rectifier instrument, 195
Recurrence paradox, 364
Red giant, 50, 652
Red shift, 299, 448
Red Spot, Jupiter, **539**
Redhead gauge, 748
Reduction, 346, 580
Reduction (electrochemistry), 201
Reduction (photosynthesis), 509
Re-entry, 646
Reeves, Hubert, **652**
Reference standards, electrical, 193
Reflectance, 516, **611**
 thin film, 722
Reflecting microscope, 481
REFLECTION, 370, 485, **611**
 coefficient of, **486**
 films, 723
 sound, 739
 sphere of, 783
 total, 486
 ultraviolet, 742
Reflection planes, Bragg, 252
Reflectivity, **611**
Reflector, microwave, 437
Reflexion. See Reflection
REFRACTION, 370, 485, 541, **614**
 angle of, **486**
 double, 360, 496, 616
 index of, 368, **486**, 495, 611
 sound, 739
Refractive index, 342, 615
Refractive index (in microscope), 481
Refractive power, magnetic lens, 213
Refractometers, 340
Refractory metals, 422
REFRIGERATION, 375, **617**
Regge poles, 264
Regge trajectories, 112
Reines, Frederick, **455**, **778**
Reissner's membrane, 310
Relative aperture, 368, 481
Relative humidity, 676
Relative permittivity, **159**
Relative stopping power, 142
Relativistic covariance, 263
Relativistic quantum field theory, 30
RELATIVITY, 262, **619**
 Doppler shift in, 176
 frequency effect, 177
 general, **620**
 in quantum theory, 573

- special, 30, **620**
- special principle of, **620**
- and time, 725
- Relaxance, 763
- RELAXATION, **622, 646, 741**
- Relaxation modulus, 762
- Relaxation time, 763
 - charge carriers, 128
 - in dielectrics, **159**
 - spin-lattice, 624
- Reluctance, 451
- Remanence, 257, **390**
- Remanent magnetization, **257**
- Remote cutoff tube, **226**
- Renormalization, 570
- Reproduction factor, 478
- REPRODUCTION OF SOUND, **625**
- Repulsion, electrostatic, 469
- Rescigno, A., 408
- Residual interaction, 475
- Residual nucleus, **470**
- Residual resistivity, **350**
- Resistance, electrical, **198, 649**
- Resistance, electrical, at high pressure, 553
- Resistance, electrical, superconductivity, **685, 686**
- Resistance, electrolytic, **346**
- Resistance, measurement of, **193**
- Resistance, plate, **225**
- Resistance thermometer, **143, 711**
- Resistanceless flow, **687**
- Resistivity, **127, 641**
- Resistivity, residual, **350**
- Resistivity tensor, 304
- Resolution, electron microscope, 211
- Resolution, microscope, **481**
- Resolving power, 429, **481**
- RESONANCE, 365, 457, 624, **626**
 - alternating current, 24
 - atomic, 3
 - cyclotron, **158**
 - diamagnetic, **158**
 - electric, 52
 - electron spin, **220, 533, 625, 672**
 - gyromagnetic, **254**
 - magnetic, **386**
 - nuclear magnetic, 169, 533
 - paramagnetic, 502
- Resonance cavity, 778
- Resonance energy, **82**
- Resonance fluorescence, **381, 447**
- Resonance frequency, 628
- Resonance neutron, 456
- Resonance radiation, 483
- Resonance transfer, 525
- Resonances, **231**
- Resonant circuit maser, **403**
- Restrahlen band, 611
- Resultant, 680
- Resultant force, 753
- Retardance, 763
- Retardation time, 763
- Retina, 766
- Reverberation, sonar, 666
- Reverberation time, **35**
- Reversibility paradox, **364**
- Reversible cell, 74
- Reversible process, 92
- Reynolds, J. A., **288**
- Reynolds, O., **280, 764**
- Reynolds number, 94, 280, 317, 392, **764**
- RHEOLOGY, **630, 765**
- Rheostat, 183
- Rho resonance, 232
- Rhodopsin, **115**
- Rhombic antenna, 27
- Richards, James A., Jr., **86, 251**
- Richardson, O. W., **224, 714**
- Richardson equation, 209, **714**
- Richardson-Dushman equation, **224**
- Richter scale, 638
- Rickets, 744
- Riemannian tensor, 408
- Righi, Augusto, **306**
- Righi-Leduc effect, **306**
- Right-hand rule, **328**
- Rigid body, **178**
- Rigid body dynamics, **178**
- Kindler, W., **622**
- Risetime, 561
- Ritter, Johann Wilhelm, **742**
- Ritz combination principle, **60**
- Robertson, B. L., **184**
- Robins, Benjamin, **67**
- Robson, J. M., **456**
- Rochelle salt, 255
- Rocket, electric, **185**
- Rocket, nuclear, **278**
- Rocket, solid propellant, **273**
- Rocket engine, 270
- Rocket engine, photon, **278**
- Rocket propulsion, 47
- Rod, vibration of, **761**
- Rod (eye), 767
- Rodgers, 715
- Roentgen, Wilhelm Konrad, 78, 306, 323, 325, 420
- Roentgen (unit), 421
- Roentgen ray, **784. See also X-ray**
- Romain, Jacques E., **726**
- Römer, Ole, **756**
- Röntgen, Wilhelm Conrad, **784. See also Roentgen**
- Room constant (acoustic), **33**
- Root-mean-square value (ac), **22, 195**
- Roozeboom, H. W. B., **506**
- Rosa, Edward Bennett, **757**
- Rosenblum, S., 468
- Rossby, C.-G., 295
- Rossby wave, 424
- Roston, S., 406
- Rotary motion, **631**
- Rotary pump, 745
- Rotation, 419
 - antiferromagnetic, **249**
 - paramagnetic, **249**
- ROTATION—CIRCULAR MOTION, **631**
- Rotational fine structure, **70**
- Rotational motion, molecule, 292
- Rotational spectra, 68
- Rothman, Milton A., **131**
- Rotor, **183, 451**
- Rouse, Arthur G., **633**
- "Rover," 265
- Row matrix, 409
- Rubber, physics of, 547
- Ruby laser, 366
- Rules, selection, **131**
- Rumford, Count, (Benjamin Thompson), **312**
- Runge's rule, **790**
- Rupp, A. F., **355**
- Rusk, Rogers D., **469**
- Ruska, E., **212**
- Russ, S., **307**
- Russell pycnometer, **157**
- Rutherford, Lord Ernest, **58, 150, 233, 466, 468, 469, 476, 771, 778**
 - atomic nucleus, 360
- Rutherford scattering experiment, 559

- Rydberg constant, 62, 134
 Ryschkewitsch, G. E., 83
- S-matrix theory**, 264
S-state wave functions, 484
 Saddington, K., 169
 Sadler, C. A., 63
 Sail, solar, 278
 Sampling theorem, 249
 Sanderson, R. T., 506
 Sanger, E. F., 708
 Sargent, W. L. W., 140
 Satellite, 47, 360
 energy of, 36, 46
 orbit, 298
 period of, 37
 Satellite (weather), 423
 Satellite tracking, 294
 Saturation magnetization, ferrimagnetism, 254
 Saturn, 535, 539
 Sawtooth wave, 454
 Scala tympani, 310
 Scala vestibuli, 310
 Scalar, 751
 Scalar field, 753
 Scale, atomic mass, 438
 Scale height (atmosphere), 538
 Scanning microscope, 215
 Scattering, 109, 704, 730
 charge carrier, 305
 fluorescent, 483
 light, 372, 484
 neutron, 456, 461
 phonon, 508
 x-ray, 120
 Scattering cross section, 109
 Scattering length, 461
 Scattering matrix, 111
 Scattering power, 461
 Scattering rate, charge carriers, 128
 Scattering theory, 111
 Schawlow, A. L., 365
 Schiff, Leonard I., 85, 415
 Schilling effusion, 157
 Schlicke, H. M., 91
 Schottky, Walter, 315
 Schottky anomaly, 315
 Schreiffer, J. R., 686
 Schrödinger, Erwin, 58, 209, 210, 216, 634, 771
 Doppler effect, 176
 SCHRÖDINGER EQUATION, 69, 251, 262, 412, 531, 572, 634, 736, 771
 Schrödinger theory (spectra), 60
 Schuler oscillation, 335
 Schumann, Victor, 741
 Schumann plate, 742
 Schumann region, 741
 Schuster, Sir Arthur, 208
 Schwarzschild, 622
 Schwarzschild solution (gravitation), 300
 Schwinger, J., 109, 111, 557
 Scintillation counter, 459, 522
 Scintillation detector, 465, 466
 Screen grid, 225
 Screw, machine, 647, 648
 Screw dislocation, 145
 Screw orientation, 127
 Seaborg, Glenn T., 736
 Sears, G. W., 145
 Second, 132
 atomic, 54
 Second law of thermodynamics, 92
 Second quantization, 263
- Secondary cell, 74
 SECONDARY EMISSION, 258, 519, 635
 Secondary ionization, 190
 Secondary particle, 465
 Secondary radiation, cosmic rays, 138
 Sector-focused cyclotron, 12
 Sedimentation constant, 94
 Sedimentology, 295
 Seebeck, T. J., 720
 Seebeck coefficient, 720
 Seebeck effect, 720
 Seeding, cloud, 106, 425
 Segrè, G., 408
 Seidman, Arthur H., 173
 Seismic wave, 638
 Seismograph, 638
 SEISMOLOGY, 296, 344, 638
 Selection rule, 131, 237, 671
 band spectroscopy, 69
 spectra, 61
 Selective radiation, 588
 Self-capacitance, 331
 Self-consistent approximation, 540
 Self-consistent field method, 399
 Self-diffusion, 170
 surface, 690
 Self-inductance, 330
 Sellmeier, W., 4
 Sellmeier equation, 4, 615
 SEMICONDUCTOR, 239, 251, 252, 304, 513, 639, 662
 cyclotron resonance in, 154
 extrinsic, 129
 film, 723
 intrinsic, 129
 laser, 366
 n-type, 129
 p-type, 129
 resistivity, 128
 surface of, 692
 Semiconductor (transistor), 728
 Semiconductor rectifier, 608
 Semiconductor-radiation detector, 465, 467
 Semi-metal, 720
 Semipermeable substance, 500
 Semipolar bond, 444
 Sengers, J. M. H. Levelt, 119
 Sensitized fluorescence, 381
 Sensors, 77
 Separative power (centrifuge), 95
 Separative work (centrifuge), 95
 Series, electromotive, 200
 Series, fundamental (spectra), 61
 Series, principal (spectra), 61
 Series, sharp (spectra), 61
 Series inductance, 331
 Serson, 301
 SERVOMECHANISM, 642
 Shadow microscope, 215
 Shaknov, I., 122
 Shankland, R. S., 429
 Shannon, Claude E., 149, 239, 409
 Sharkey, A. G., Jr., 406
 Sharp series (spectra), 61
 Shaw, E. A. G., 200
 Shear, 630
 Shear force, 281
 Shear wave, 739
 Sheehan, William F., 99
 Shell, closed, 61
 Shell, electron, 237
 Shell, quantum, 502
 Shell model, nucleus, 399, 474, 560
 Sheppard, C. W., 408

- Shielding, electrostatic, **417**
 Shielding, magnetic, **417**
 Shielding, reactor, 473, **603**
 Shift, isotope, **62**
 Shift, parallactic, **42**
 Shire, E. S., **198**
 Shock front, **280**
 Shock tube, **646**
 SHOCK WAVES, 93, 346, **645**
 Shoran, **758**
 Short range force, **477**
 Shot noise, **418**
 photomultiplier, **522**
 Shottky, W., **715**
 Shower, cosmic ray, **138**
 Shubnikow, **502**
 Shugart, Howard A., **53**
 Shurcliff, William A., **545**
 Sidran, Miriam, **515**
 Siegbahn, M., **785**
 Sievert, R., **307**
 Sigma hyperon, **229**
 Signal, oscilloscope, **497**
 Sikkeland, T., **736**
 Silicon, in transistors, **728**
 Silicon cell, **657**
 Silsbee rule, **685**
 Silver halide, 514, **515**
 Simon, F., **142**
 Simple harmonic motion, **69**, 454, **760**
 SIMPLE MACHINES, **647**
 Sine galvanometer, **396**
 Sinelnikov experiment, **287**
 Singlet state, 61, **382**
 Sitter, Willem de, **140**
 Skaggs, Lester S., **421**
 SKIN EFFECT, **649**
 anomalous, **253**
 Skin friction drag, **21**
 Skolnik, Merrill L., **577**
 Sky waves, **436**
 Sleator, W. W., **68**
 Slepian, J., **637**
 Sliding (friction), **284**
 Slifkin, Lawrence, **171**
 Slotted line, **777**
 Slow neutron, **456**
 Smekal, A., **599**
 Smolt, J., **625**
 Smith, James T., **388**
 Smog, **743**
 Smoluchowski, R., **373**
 Smyth, Charles P., **160**, **542**
 Snap diode, **563**
 Snell, Willebrord, **612**, 614, **739**
 Snell's law, 370, 496, 557, **612**
 Snoek, J. L., **254**, **741**
 Snoek peaks, **741**
 Snow blindness, **743**
 Snyder, **704**
 Soddy, Frederick, **352**
 Sodium light, **371**
 Solar activity, **660**
 Solar battery, **514**
 Solar burst, **592**
 Solar cells, 186, 511, 653, **656**
 Solar corona, **395**
 Solar eclipse, **660**
 Solar energy, 185, 470, **659**
 SOLAR ENERGY SOURCES, **650**
 SOLAR ENERGY UTILIZATION, **653**
 Solar flare, 138, 395, **661**
 Solar furnace, **589**
 SOLAR PHYSICS, 393, **659**
 Solar plasma, **667**
 Solar radio wave, **592**
 Solar sail, **278**
 Solar spectrum, **744**
 Solar system, 41, **50**, **398**
 Solar ultraviolet, **743**
 Solar wind, **667**
 Solenoid, 332, **395**
 Solid, **674**
 density, **157**
 Solid phase, **506**
 Solid propellant, **273**
 Solid solution, **507**
 SOLID-STATE PHYSICS, 535, **661**
 SOLID-STATE THEORY, **663**
 Solidus, **507**
 Solution, heat of, **750**
 Sommerfeld, Arnold, **422**
 electron orbits, **422**
 heat capacity, **315**
 Sommerfeld theory, metallic conduction, **715**
 SONAR, **664**, **740**
 Sone, **463**
 Soo, S. L., **242**
 Soroka, Walter W., **18**
 Sorption pump, **747**
 Sound, **16**, **31**, **534**
 absorption coefficient, **33**
 acoustic properties, **309**
 intensity, **32**
 propagation of, **399**
 reproduction of, **625**
 velocity, **118**, **739**
 Sound navigation and ranging, **664**
 Sound power, **32**
 Sound pressure, **310**, **453**, **463**
 Sound wave, **280**, **528**, **739**
 Space, curved, **299**
 Space, Euclidean, **139**
 Space, physics of, **666**
 Space charge, **192**, **224**
 electrostatic thrustors, **188**
 surface, **692**
 Space flight, **36**
 Space inversion, **229**
 SPACE PHYSICS, **666**
 Space probe, **668**
 Space time, **621**, **622**, **726**
 Space vehicle, **186**
 temperature control, **588**
 Space wave, **436**
 Spark breakdown, **191**
 Spark chamber, **465**, **467**
 Spatial memory, **301**
 Special principle of relativity, **620**
 Special relativity, **30**, **620**
 Special unitary group, **229**, **231**
 Specific activity, **355**
 Specific conductance (dielectrics), **160**
 Specific fuel consumption, **271**
 Specific gravity, **156**, **157**
 Specific heat, **244**, **250**, **291**, **312**, **313**, **508**
 Specific heat capacity, **311**
 Specific impulse, **47**, **185**, **271**
 Specific inductive capacity, **159**
 Specific mass, **184**
 Spectacle lens, **367**
 Spectra, **60**, **238**, **669**
 absorption, **2**
 atomic, **59**, **668**
 electromagnetic, **669**, **741**
 emission, **371**
 line, **59**
 molecular, **68**, **337**

- Spectra** (*Cont.*)
 phonon, 508
 Raman, 599, 669
 rotational, 68
Spectral concentrations, 113
Spectral emissivity, 563
Spectral line, radio astronomy, 593
Spectral radiance, 563
Spectrograph, 6
 mass, 404
Spectrometer, 495
 mass, 404
 radiometric, 590
Spectrometry, 459
Spectrophotometry, 671
 absorption, 6
 infrared, 339
Spectroscope, 6, 495
Spectroscopic fine structure, 219
Spectroscopic splitting factor, 248
SPECTROSCOPY, 371, 668, 713
 Auger, 63
 band, 68
 fields of, 670
 infrared, 339
 microwave, 433
 molecular, 68
Specular reflectance, 613
Sperry, Elmer, 301
Sphere of reflection, 783
Spherical aberration, 368
Spherical surface, 491
Spherical top molecule, 70
Spike, thermal, 350
Spin, 456
 electron, 219
 isotopic, 231
 nuclear, 52, 62
 particle, 684
 photon, 522
Spin angular momentum, 237, 256
Spin exchange, 484
Spin magnetic quantum number, 503
Spin moments, 175
Spin orientation, particles, 111
Spin quantum number, 389
Spin temperature, 388
Spin wave, 256
Spin-lattice relaxation time, 624
Spin-orbit coupling, 61, 142
Spin-spin interaction, 388, 402, 672
Spinks, J. W. T., 581
Spinor analysis, 409
Spiral growth, dislocation, 664
Splitting factor, spectroscopic, 220, 248
Sporadic E, 349
Spread F, 349
Spur gear, 649
Sputnik, 360
Sputter-ion pump, 747
Square matrix, 410
Square wave, 22, 454, 498
Stabilizer, ship, 301
Standard cell, 74, 194
Standard deviation, 134
Standard potential, electrode, 202
Standards, reference, electrical, 193
Stanton number, 317
Stapedius, 310
Stapes, 309
Star, 41, 48, 139
 diameter, 108
 distance of, 43
 evolution of, 50
 exploding, 137
 fixed, 178
 giant, 50
 main sequence, 50
 mass of, 49
 motion of, 41
 radio, 592
 spectra of, 49
Star catalogues, 41
Stark, Johannes, 238, 792
Stark effect, 62, 83, 174, 238, 402, 433, 542, 790, 792
State, change of, 675
State, equation of, 289
State, excited, 117
State vector, 567
STATES OF MATTER, 674
STATIC ELECTRICITY, 197, 677
Static friction coefficient, 283
Static multiplier, 521
STATICS, 679
 fluid, 281
Station keeping, 39
Statistical equilibrium, 241
Statistical lag, 191
STATISTICAL MECHANICS, 98, 251, 399, 432, 680, 712
 quantum, 682
Statistical thermodynamics, 719
Stator, 183, 451
Steam plant, nuclear, 184
Steam turbine, 183
Stearns, Robert L., 84
Stefan, Josef, 319, 586
Stefan-Boltzmann constant, 134, 319
Stefan-Boltzmann law, 586
Stellar distance determination, 43
Stellar evolution, 50, 652
Stellar spectra, 49
Stellarator experiment, 286
Steno, 147
Step function, 498
Stephens, William E., 221
Stephenson, Reginald J., 420
Stereoisomerism, 71
Stereophonic sound, 626
Stereoregularity, 547
Stereoscopic depth, 768
Stereoscopic microscope, 433
Sterilization of food, 472
Stern, Otto, 51, 52
Stern-Gerlach experiment, 51, 219
Sternglass, E. J., 637
Stevenson, J. S., 408
Stevin (or Stevinus), Simon, 679
Steward, G. C., 1
Stibitz, George, 124
Stick-slip, 284
Stilb, 379
Stilwell, A. R., 178
Stirling engine, 653
Stoichiometry, 439
Stokes, George Gabriel, 280, 367, 742, 764, 765
Stokes' emission, 381
Stokes' law, 94, 209, 280, 765
Stokes line, 367, 599
Stokes shift, 117
Stokes' theorem, 294
Stokes (unit), 764
Stoney, G. Johnstone, 208
Stop, aperture, 368
Stop, lens, 482
Stopping number, 583

- Stopping power, 142
 relative, 142
 mass, 583
 Storage battery, 72
 Storage modulus, 762
 Storage rings, 12
 Storms, magnetic, 65, 344
 Strain, 512
 viscoelasticity, 762
 Strain dyadic, 181
 Strain energy, 183
 Strand, K. Aa., 44
 Strange particle, 229
 Strangeness, conservation of, 131
 Stranski, I. N., 127
 Strapdown system, guidance, 335
 Strassmann, F., 264
 Stratosphere, 296
 Stratospheric fallout, 247
 Streaming potential, 202
 Street, Kenneth, Jr., 735
 Stress, 512
 internal, 181
 viscoelasticity, 762
 Stress corrosion, 423
 Stress dyadic, 181
 String, in wave motion, 775
 String, vibration of, 760
 Stripping (ionization), 347
 Stripping reaction, 471
 Strong coupling, 346
 STRONG INTERACTIONS, 29, 228, 229, 570, 683
 Strong nuclear force, 683
 Strontium, 57
 Structure, crystal, 147
 Structure, hyperfine, 62
 Structure, nuclear, 474
 Structure factor, 784
 Stuhlinger, E., 188
 SU_2 and SU_3 , 231
 Sublimation, 105, 662, 676
 heat of, 750
 Sublimation energy, 692
 Sub-shell, closed, 61
 Subshell, electron, 237
 Substage condenser, 483
 Suchy, K., 557
 Sudarshan, E. C. G., 233, 455
 Sun, 45, 344, 357, 651, 653, 659
 International Years of the Quiet, 344
 ionized gas from, 138
 luminosity, 659
 luminosity unit, 49
 particles from, 667
 physics of, 659
 velocity of, 42
 Suna, Andris, 243
 Sunspots, 344, 393, 660
 Superconducting magnet, 144
 Superconducting ring, 687
 SUPERCONDUCTIVITY, 143, 400, 685, 687, 688, 737, 741
 Supercooling, 105, 377
 Superexchange, 28, 390
 SUPERFLUIDITY, 143, 377, 400, 687
 Superheating, 377
 Supernova, 137
 Superparamagnetism, 391
 Superposition principle, Boltzmann, 280, 762
 Supersaturation ratio, 126
 Supersonic aircraft, 646
 Supersonic flow, 280
 Suppressor grid, 225
 Surface, aspheric, 367
 Surface, diffuse, 588
 Surface, structure of, 690
 Surface defect, 661
 SURFACE PHYSICS, 689
 Surface self-diffusion, 690
 Surface space charge, 692
 SURFACE TENSION, 676, 689, 692
 Surface tension wave, 776
 Surface wave, 775
 Surveying, 294
 Susceptibility, 175, 249, 389, 501
 electric, 205
 Krishnan's method, 175
 magnetic, 155, 158, 206
 static (antiferromagnetism), 28
 Susskind, Charles, 227
 Sutton, George P., 278
 Sutton, Richard M., 360
 Svedberg, The (in full, Theodore), 94
 Svedberg (unit), 94
 Sverdrup, H. U., 295
 Swenson, C. A., 553
 Swift, J. D., 193
 Switch, 254
 SYMBOLS, UNITS AND NOMENCLATURE IN PHYSICS, 694
 Symmetric top, 434
 Symmetry, 29, 110, 130, 713
 crystal, 147
 internal, 231
 mirror, 131
 unitary, 231
 Symmetry point group, 445
 Synapse, 408
 Synchrocyclotron, 10, 152
 Synchronous motors, 452
 SYNCHROTRON, 11, 150, 469, 704
 electron, 11
 proton, 11
 Synchrotron radiation, 137
 Syndiotactic polymer, 547
 Synodic periods, 39
 Systematic error, 415
 Szilard, L., 355, 478
 Szilard-Chalmers process, 355

 t-matrix, 400
 T-F emission, 258
 Tait, Peter G., 751
 Tamm, Ig., 97, 692
 Tamm state, 692
 Tamm-Dancoff calculation, 546
 Tangential force, 281
 Taylor, L. S., 307, 420
 Teaney, Dale T., 29
 Tee, waveguide, 777
 "Teledeltos," 708
 TELEGRAPHY, 707
 TELEMETRY, 423, 708
 Teleprinter, 707
 Telescope, 480
 Telescope, Mount Palomar, 482
 Telescope, Yerkes, 482
 Telescope objective, 369
 Television, 115
 Tellurometer, Wadley's, 759
 Temperature, 91, 311, 675, 709
 AND THERMOMETRY, 709
 critical, 749
 Curie, 256
 earth, interior, 296
 gas, 291
 glass transition, 547

- Temperature measurement, **563**
 Temperature scale, 240, 311, 710
 International Practical, 564
 Temperature, transition, superconductivity, 685
 Tensile force, **281**
 Tensile stress (in liquid), **93**
 Tension, surface, 689, **692**
 Tensor tympani, 310
 Teorell, Torsten, 407
 Tepley, Norman, **254**
 Terbium earths, 601
 Terminal impedance, **25**
 Ternary fission, **268**
 Terrell, James, **269**
 Tesla, N., 451
 Testing, ultrasonic, **740**
 Tetrode, **223, 225**
 Thales of Miletus, 233, **323**
 Thekaekara, Matthew P., **343**
 THEORETICAL PHYSICS, 408, 535, **712**
 Theory of vision, Hering, 116
 Thermal conductivity, **315**
 Thermal diffusion, 316, **362**
 Thermal diffusion (isotope), **354**
 Thermal effect (measurement), **417**
 Thermal electron, 606
 Thermal emission, 337, 520
 Thermal energy, 507
 Thermal expansion, **244, 314**
 Thermal fluctuation, 418
 Thermal neutron, **456, 458, 473** (footnote), 478
 Thermal radiation, 570, **585**
 Thermal spike, **350**
 Thermionic cell, **186**
 Thermionic emission, 258
 Thermionic emission microscope, **215**
 THERMIONICS, **714**
 Thermochemical calorie, **133**
 Thermocouple, 338, 590, 710, **721**
 Thermocouple gauge, **748**
 Thermocouple instruments, **195**
 Thermodynamic equilibrium, **289**
 Thermodynamic potentials, **718**
 Thermodynamic process, 91
 Thermodynamic state, 91
 THERMODYNAMICS, 240, 244, 375, 532, 533, 682, 711, **716**
 galvanic cell, **73**
 phase rule, **506**
 second law of, 241
 zeroth law, **240**
 Thermoelectric effect, 711
 Thermoelectric power, 720
 Thermoelectric refrigerator, **618**
 Thermoelectric voltage, 306
 THERMOELECTRICITY, 618, **720**
 Themography, 590
 Thermomagnetic effect, 306, 619
 Thermomagnetic refrigerator, 619
 Thermometer, 241, 311, **711**
 clinical, 711
 gas, **143, 711**
 maximum, 711
 minimum, 711
 resistance, **143**
 Thermometry, **563, 709, 711**
 radiation, **563**
 Thermonuclear device, **265, 479**
 Thermonuclear reaction, 56, 651
 Thermonuclear weapon, 56
 Thermopile, 590, **721**
 Thermostat, 249, 711
 Thermostatics, laws of classical, **240**
 "Theta" pinch experiment, **286**
 THIN FILM, **721**
 electron diffraction by, 210
 magnetic, **391**
 superconductivity in, 685
 thermionic properties, 715
 Thirring, W., 112
 Thixotropic substance, 765
 Thomas, L. H., **153**
 Thompson, Barbara A., **539**
 Thompson, Benjamin, (Count Rumford), 312
 Thompson, Stanley G., 735
 Thomson, Elihu, 451
 Thomson, G. P., 209, 210, 531
 Thomson, Sir Joseph John, 57, 120, 208, **323, 325, 346, 404, 512, 635, 714, 784**
 positive ray, 352
 protons, 559
 Thomson, William, (Lord Kelvin, *q.v.*), 92, 142, 291
 Joule-Thomson expansion, 142
 Thomson cross section, **134**
 Thomson effect, **720**
 Thorium, 472
 Thorsen, A. C., **156**
 Three-body problem, 474
 Thrust, **21, 47, 271**
 Thrust chamber, rocket, 272
 Thrust vector control, **275**
 Thrustor, **184, 188**
 Thrust-weight ratio, 185
 Thun, Rudolf, F., **724**
 Thyatron, 606, **609**
 Thyristor, 606
 Tides, **296**
 Timbre, **454**
 TIME, **724**
 astronomical, **725**
 relaxation, 388, **622**
 Time constant, capacitor, **89**
 Time constant, ionization chamber, 466
 Time marker, 561
 Time reversal, 229
 Time reversal invariance, **131**
 Time-base generator, 496
 Time-temperature superposition, **762**
 Titan (satellite), 539
 Todd, Paul W., **585**
 Tone color, 454
 Top, symmetric, 434
 Toroid, 332
 Torque, **327, 633, 680**
 gyroscope, **301**
 Torque motors, 183
 Torr, 745
 Torsional wave, **775**
 Total radiation pyrometer, 564
 Tourmaline, 496
 Tousey, Richard, **744**
 Townes, C. H., 365, **400**
 "Toy Top," **286**
 Trace analysis, 458
 Tracers, radioactive, **594**
Traité de Chimie, **233**
 Trajectory, **67**
 powered, 48
 Transducer, 199, **305, 739**
 ferroelectric, 739
 magnetostrictive, 739
 piezoelectric, 739
 Transduction, **199**
 Transfer, resonance, 525
 Transfer function, servomechanism, 643
 Transfer orbit, 47
 Transfer reaction, 471

- Transformation, gauge, **565**
Transformation equations, Galilean, **621**
Transformation of coordinates, **399**
TRANSFORMER, **75, 726**
 current, **726**
 emf, **328**
 instrument, **196**
 narrow-band, **727**
 pulse, **727**
 voltage, **196**
 wide-band, **727**
Transient phenomena, **418**
TRANSISTOR, **639, 642, 728**
 avalanche, **562**
 junction, **728**
 planar, **729**
 point-contact, **728**
Transition, electronic, **669**
Transition, recoilless, **447**
Transition element, **504**
Transition energy levels, **238**
Transition probabilities, **484, 774**
Transition temperature, superconductivity, **685, 686**
Transmission, microwave, **436**
Transmission line, **776**
Transmittance, **516**
Transport coefficient, plasma, **683**
Transport phenomena, **407, 664**
TRANSPORT THEORY, **363, 730**
 neutron, **141**
TRANSURANIUM ELEMENTS, **732**
Transverse magnetoresistance, **305**
Transverse wave, **739, 775**
Trap, electron, **511, 611**
Trapping, Lorentz, **287**
Travel, interplanetary, **188**
Traveling-wave maser, **402**
Traveling-wave tube, **226**
Traverse, **294**
Treble boost, **626**
Tremolo, **453**
Triangulation, **294**
Triboelectric series, **678**
Tribus, M., **240**
Trichromatic vision, anomalous, **113**
Triclinic crystal, **147**
Triggered spark gap, ignitron, **562**
Trilateration, **294**
Triode, **223**
Tripack film, **515**
Triple (vectors), **752**
Triple point, **377, 507, 676, 710**
Triplet, **61**
Triplet exciton, **243**
Triplet production, **584**
Triplet state, **382**
Tristimulus values, **113, 115**
Tritium, **285, 459, 559**
Triton, **559**
Tropopause, **426**
Troposphere, **296, 426**
Tropospheric fallout, **247**
Troup, G. J., **403**
Frouton's rule, **378**
True anomaly, **38**
Trump, John G., **16, 323**
Tschoegl, N. W., **764**
Tsunami, **297**
Tsuya, **398**
Tube
 cathode ray, **217**
 density gradient, **157**
 electron, **222**
 mercury vapor, **224**
 microwave, **226**
 multifunction, **226**
 remote cutoff, **226**
 variable-mu, **226**
Tuning (electronics), **24**
Tuning fork, **454**
Tunnel diode, **562, 738**
Tunnel effect, **774**
TUNNELING, **562, 596, 641, 736**
Turbine, steam, **183**
Turbine, water, **183**
Turboelectric systems, electric propulsion, **86**
Turbojet, **278**
Turbulence, sound source, **529**
Turbulent flow, **280**
Turnover frequency, **625**
Twenty-four hour orbit, **39**
Twilight vision, **115**
Twilightglow, **65**
Two color pyrometer, **564**
Tympanic membrane, **309**
Tyndall, E. P. T., **379**
Tyndall, John, **372**
 scattering, **372**
U. S. Naval Research Laboratory, **744**
Ubaldi, Guido, **679**
Uhlenbeck, Goerge Eugene, **209, 219, 236, 790**
Ulbricht sphere, **517**
Ultracentrifuge, **95, 441, 443**
Ultrasonic power, **740**
Ultrasonic testing, **740**
ULTRASONICS, **17, 739**
 phonon-electron amplifier, **200**
Ultrasonics and phonons, **508**
Ultrastability, **249**
Ultraviolet, **70, 370**
 detection of, **742**
 laser, **366**
 solar, **743**
 "Ultraviolet Catastrophe," **586**
 "Ultraviolet Difficulty," **571**
Ultraviolet microscope, **481**
ULTRAVIOLET RADIATION, **741**
Uncertainty principle, Heisenberg, **320, 574, 713**
Unified atomic mass unit, **133**
Unipotential, lens, **213**
Unit, **132, 694**
Unit, practical system, **303**
Unit cell, crystal, **147, 442, 460, 783**
Unit matrix, **413**
Unitary group, special, **231**
Unitary symmetry, **231**
Universal Fermi interaction, **779**
Universe, expanding, **139**
Universe, models, **139**
Univibrator, **561**
Unwin, Robert S., **66**
Uranium, **55, 462, 472**
Uranus, **535**
Urbach law, **243**
Ure, Roland W., Jr., **721**
Ursell, H. D., **290, 682**
Ussing, H. H., **407**
V_K-center, **117**
V-value, **615**
Vacancy, in crystal, **171, 260, 350**
Vacuole, **500**
Vacuum, **281, 745**
Vacuum evaporation, **721**
Vacuum gauge, **748**
Vacuum pump, **745**

- VACUUM TECHNIQUES, 745**
 Vacuum tube, **221**
 Vacuum ultraviolet, **741**
 Vacuum-tube voltmeters, **195**
 Valdes, L. B., **729**
 Valence, **73, 444, 533**
 Valence band, **129, 366, 510, 640, 662**
 Valency, **444**
 Value, effective, **195**
 Van Allen, James Alfred, **138, 577**
 Van Allen radiation belt, **138, 577, 667**
 Van Alphen, P. M., **144, 155**
 Van de Graaff accelerator, **9, 12, 459**
 Van der Pol, Balth, **104**
 van der Waals, Johannes Diderik, **289**
 van der Waals equation, **289, 530**
 van der Waals forces, **18**
 van Helmont, Jan Baptista, **233**
 van't Hoff, Jacobus Hendricus, **440, 499, 530**
 van't Hoff's equation, **499**
 VanZandt, T. E., **349**
 Vapor, **674**
 Vapor pressure, **376, 442, 676**
VAPOR PRESSURE AND EVAPORATION, 749
 Vapor pump, **745**
 Vaporization, heat of, **312**
 Vaporization curve, **507**
 Vaporous, **507**
 Variable mu-tube, **226**
 Variational emf, **328**
 Vector, **751**
 polarization, **97**
 position, **179**
 Vector algebra, **751**
 Vector analysis, **409**
 Vector field, **753**
 Vector Laplacian, **551**
VECTOR PHYSICS, 751
 Vector potential, magnetic, **206**
 Vector space, **751**
 Vega, **358**
 Vegard, L., **295**
 Veksler, V., **151, 704**
 Velitschko, **759**
 Velocity, **36, 179, 326, 573**
 ballistics, **67**
 drift (charge carriers), **128**
 effective jet, **185**
 group, **776**
 phase, **776**
 projectile, **67**
 rocket, **271**
 sound, **739**
 target, **576**
 wave, **756, 775**
 Velocity asymptote, **38**
VELOCITY OF LIGHT, 755
 Velocity space instability, **287**
 Vening-Meinesz, F. A., **295**
 Venturi, **280**
 Venus, **535**
 Verneuil method, **146**
 Vestibule, **310**
 Vestuli, **309**
VIBRATION, 454, 759
 in acoustic waves, **664**
 mechanical, **416**
 Vibration band (spectra), **70**
 Vibration frequency, molecular, **600**
 Vibrato, **453**
 Vieille, Paul Marie Eugène, **646**
 Vignetting, **482**
 Villard, **467**
 Villchur, Edgar, **626**
 Vinci, Leonardo da, **324**
 Virial coefficient, **719**
 Virial equation of state, **289, 682**
 Virtual photon, **228**
 Virtual work, **680**
Vis viva energy equation, **36**
VISCOELASTICITY, 762
 Viscometer, **765**
VISCOSITY, 279, 630, 764
 in electrolytic conductivity, **201**
 polymers, **547**
 vitreous material, **769**
 Viscosity coefficient, **279, 764**
 Viscosity index, **765**
 Visibility, fringes, **107**
 Vision, Hering theory, **116**
VISION AND THE EYE, 766
 Vitamin D, **744**
 Vitreous humor, **766**
VITREOUS STATE, 769
 Vlasov, **541**
 Vlasov approximation, **540**
 Vulcanology, **296**
 Volt, **202**
 Volta, Alessandro, **78, 200**
 Voltage, **102, 329**
 effective, **22**
 measurement of, **193**
 voltage divider, **194**
 Voltage standing-wave ratio, **777**
 Voltage transformer, **196**
 Voltaic cell, **72**
 Voltaic pile, **200**
 Volterra, Vito, **406**
 Voltmeter, **193**
 electrodynamical, **196**
 vacuum tube, **195**
 Volume, gas, **291**
 von Alteneck, **450**
 von Ardenne, M., **215**
 von Aulock, Wilhelm H., **254**
 von Guericke, Otto, **323**
 von Helmholtz, Hermann Ludwig Ferdinand. *See* Helmholtz
 von Laue, Max, **147**
 von Linde, Carl, **375**
 von Neumann, John, **409, 681**
 von Neumann's equation, **681**
 von Siemens, W., **450**
 VSWR, **777**
 Wadley's tellurometer, **759**
 Wahl, A. C., **733**
 Waller, I., **164**
 Walsh, D., **591**
 Walter, W. G., **150**
 Walton, E. T. S., **479**
 Walton, J. R., **736**
 Wangsness, Roald K., **714**
 Wannier-Mott model, **242**
 Water turbine, **183**
 Water wheel, **183**
 "Water-boiler," **266**
 Waterston, J. J., **362**
 Watson, Kenneth M., **112**
 Watt, **780**
 Wattenberg A., **123**
 Wattmeter, **193, 196**
 Wave
 acoustic, **16**
 compressional, **775**
 electromagnetic, **198, 554**
 gravitational, **300**

- hemispheric long, 424
- hydro-magnetic, 348
- ion-acoustical, 348
- Lamb, 739
- longitudinal, 739, 775
- Raleigh, 739
- Rossby, 424
- sawtooth, 454
- shear, 739
- shock, 93
- sound, 16, 528, 739
- square, 22, 454
- transverse, 739, 775
- Wave equation, 262
 - Schrödinger, 262, 531, 572, 772
- Wave front, 97
- Wave function, 83, 573, 634, 736, 772
 - S-state, 484
- WAVE MECHANICS, 236, 321, 412, 477, 530, 771
- WAVE MOTION, 634, 775
- Wave number, 59, 69, 337
- Wave packet, 574
- Wave propagation, 557
- Wave pulse, 739
- Wave train, 494
- Wave vector, 154
- Wave vector space, 154
- Wave velocity, 756
- WAVEGUIDE, 403, 436, 776
- Waveguide (accelerators), 10
- Waveguide mode, TM_{01} , 6
- Wavelength, 775
 - electron, 211
 - Compton, 121
 - light, 370
- Wavelength and resolution, 481
- WEAK INTERACTION, 228, 455, 683, 778
- Weapon, thermonuclear, 56
- Wear, 284
- Weather, 297, 423
 - Coriolis effect on, 136
- Weather modification, possibility of, 106
- Weaver, Elbert C., 158
- Webb, M. B., 692
- Weber, J., 300
- Weber, Wilhelm, 757
- Weber (unit), 328
- Wedge, 647
- Weeks, W. L., 208
- Wehnelt, A., 715
- Weight, 44, 298, 404
 - chemical equivalent, 201
 - electrochemical equivalent, 346
 - molecular, 439, 445
- Weightlessness, 44, 79
- Weinberg, S., 111
- Weiss, Pierre, 390, 501
- Weiss law, 246
- Welsbach mantle, 588
- Welsby, Vernon G., 332
- Welsh, H. L., 601
- Wentzel, Kremers, Brillouin approximation, 736
- Wert, Charles A., 663
- Weston cell, 74, 194
- Westphal balance, 157
- Whatley, Linda S., 672
- Wheatstone, C., 450
- Wheatstone bridge, 194
- Wheatstone network, 194
- Wheel, 647
- Wheeler, Gershon J., 778
- Wheeler, J. A., 264
- Whisker, 145, 664
- White, Milton G., 706
- White, R. S., 578
- White dwarf, 50, 652
- White light, 370
- White noise, 453
- Whitehead, A. N., 409
- Wide-band transformer, 727
- Wien, Wilhelm, 319, 404
- Wien, displacement law, 319, 586, 587
- Wiener, Norbert, 149, 249, 409
- Wiener-Hopf equation, 149
- Wiener-Khinchine theorem, 149
- Wigner, Eugene P., 112, 409
- Wigner-Seitz cell model, 399
- Williams, Ferd, 384
- Wills, John H., 507
- Wilson, A. J. C., 784
- Wilson, Charles Thompson Rees, 63, 126
- Wind, 136
- Wind tunnel, 21
- Wind velocity, geostrophic, 136
- Window, atmospheric, 338
- Window, oval, 310
- Window, ultraviolet, 742
- Wing, G. Milton, 732
- WKB approximation, 736
- Wolf, E., 107
- Wolfe, Hugh C., 703
- Wollan, E. O., 307
- Wollaston, William Hyde, 3, 544
- Womersley, J., 406
- Wood, Elizabeth A., 690
- Wood, R. W., 258, 599
- Woodbury, E. J., 366
- Woods, Joseph F., 511
- Woods, R. J., 581
- Work, 87, 327, 753
 - in machines, 647
- WORK, POWER AND ENERGY, 780
- Work, thermodynamics, 718
- Work, virtual, 680
- Work function, 223, 513, 690, 715
- Work hardening, 422
- Working distance, 431
- World models, 139
- Worm and wheel, 649
- Wow, 626
- Wright, Sewall, 406
- Wu, C. S., 122
- Xenon compounds, 102
- "Xerography," 515
- X_1 hyperon, 229
- X_1 resonance, 232
- X-RAY, 63, 120, 306, 420, 461, 581, 741, 782, 784
- X-ray, characteristic, 238, 354
- X-ray, in cosmic rays, 137
- X-ray, spectrum, 669
- X-ray absorption spectra, 5
- X-ray crystallography, 147
- X-RAY DIFFRACTION, 147, 442, 782
- X-ray diffraction (polymers), 547
- X-ray fluorescence, 669
- XUV, radiation, 741
- Y resonance, 232
- Yagi antenna, 27
- Yang, C. N., 131, 455, 779
- Yarwood, John, 749
- Year, International Geophysical, 344
- Yeh, Hsuan, 327
- Yerkes telescope, 482

Young, Thomas, 106, 182, **340**, 426
 interference, **340**, **494**
 polarization, 543
 sound recording, 625

Young's modulus, 182

Young-Helmholtz theory, 116

Yttrium earths, 601

Yukawa, Hideka, **263**, 478, 560, 779

"Z" pinch experiment, **286**

Zarem, A. M., **658**

Zeeman, Pieter, 57, 208, 209, 238, 393, **790**

ZEEMAN AND STARK EFFECTS, 790

Zeeman effect, 62, 209, 220, 238, 402, 450, **790**

Zeeman resonance, 485

Zener breakdown, 738

Zener tunneling, 738

Zernike, F., **1**, 164, **432**

Zeroth law of thermodynamics, **240**, **717**

Zienau, S., 545

Zimm, 548

Zinn, W. H., 478

Zone melting, **146**

Zucker, Alexander, **471**

Zworykin, V. K., **215**